Exhibit 61

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Moving to a World Beyond "$p < 0.05$"

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "p<0.05", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

To link to this article: https://doi.org/10.1080/00031305.2019.1583913

Published online: 20 Mar 2019.

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913

EDITORIAL

Moving to a World Beyond "$p < 0.05$"

Some of you exploring this special issue of *The American Statistician* might be wondering if it's a scolding from pedantic statisticians lecturing you about what *not* to do with p-values, without offering any real ideas of what *to do* about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.

1. "Don't" Is Not Enough

There's not much we can say here about the perils of p-values and signiffcance testing that hasn't been said already for decades (Ziliak and McCloskey 2008; Hubbard 2016). If you're just arriving to the debate, here's a sampling of what not to do:

- Don't base your conclusions solely on whether an association or effect was found to be "statistically signiffcant" (i.e., the p-value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically signiffcant.
- Don't believe that an association or effect is absent just because it was not statistically signiffcant.
- Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientiffc or practical importance based on statistical signiffcance (or lack thereof).

Don't. Don't. Just…don't. Yes, we talk a lot about don'ts. The *ASA Statement on p-Values and Statistical Significance* (Wasserstein and Lazar 2016) was developed primarily because affer decades, warnings about the don'ts had gone mostly unheeded. The statement was about what not to do, because there is widespread agreement about the don'ts.

Knowing what not to do with p-values is indeed necessary, but it does not suff ce. It is as though statisticians were asking users of statistics to tear out the beams and struts holding up the ediffce of modern scientiffc research without offering solid construction materials to replace them. Pointing out old, rotting timbers was a good start, but now we need more.

Recognizing this, in October 2017, the American Statistical Association (ASA) held the Symposium on Statistical Inference, a two-day gathering that laid the foundations for this special issue of *The American Statistician*. Authors were explicitly instructed to develop papers for the variety of audiences interested in these topics. If you use statistics in research, business, or policymaking but are not a statistician, these articles were indeed written with YOU in mind. And if you are a statistician, there is still much here for you as well.

The papers in this issue propose many new ideas, ideas that in our determination as editors merited publication to enable broader consideration and debate. The ideas in this editorial are likewise open to debate. They are our own attempt to distill the wisdom of the many voices in this issue into an essence of good statistical practice as we currently see it: some do's for teaching, doing research, and informing decisions.

Yet the voices in the 43 papers in this issue do not sing as one. At times in this editorial and the papers you'll hear deep dissonance, the echoes of "statistics wars" still simmering today (Mayo 2018). At other times you'll hear melodies wrapping in a rich counterpoint that may herald an increasingly harmonious new era of statistics. To us, these are all the sounds of statistical inference in the 21st century, the sounds of a world learning to venture beyond "$p < 0.05$."

This is a world where researchers are free to treat "$p = 0.051$" and "$p = 0.049$" as not being categorically different, where authors no longer ffnd themselves constrained to selectively publish their results based on a single magic number. In this world, where studies with "$p < 0.05$" and studies with "$p > 0.05$" are not automatically in conffict, researchers will see their results more easily replicated—and, even when not, they will better understand *why*. As we venture down this path, we will begin to see fewer false alarms, fewer overlooked discoveries, and the development of more customized statistical strategies. Researchers will be free to communicate all their ffndings in all their glorious uncertainty, knowing their work is to be judged by the quality and effective communication of their science, and not by their p-values. As "statistical signiffcance" is used less, statistical thinking will be used more.

The *ASA Statement on P-Values and Statistical Significance* started moving us toward this world. As of the date of publication of this special issue, the statement has been viewed over 294,000 times and cited over 1700 times—an average of about 11 citations per week since its release. Now we must go further. That's what this special issue of *The American Statistician* sets out to do.

To get to the do's, though, we must begin with one more don't.

2. Don't Say "Statistically Significant"

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical signifficance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically signifficant" entirely. Nor should variants such as "signifficantly different," "$p < 0.05$," and "nonsignifficant" survive, whether expressed in words, by asterisks in a table, or in some other way.

Regardless of whether it was ever useful, a declaration of "statistical signifficance" has today become meaningless. Made broadly known by Fisher's use of the phrase (1925), Edgeworth's (1885) original intention for statistical signifficance was simply as a tool to indicate when a result warrants further scrutiny. But that idea has been irretrievably lost. Statistical signifficance was never meant to imply scientifc importance, and the confusion of the two was decried soon affer its widespread use (Boring 1919). Yet a full century later the confusion persists.

And so the tool has become the tyrant. The problem is not simply use of the word "signifficant," although the statistical and ordinary language meanings of the word are indeed now hopelessly confused (Ghose 2013); the term should be avoided for that reason alone. The problem is a larger one, however: using bright-line rules for justifying scientifc claims or conclusions can lead to erroneous beliefs and poor decision making (ASA statement, Principle 3). A label of statistical signifficance adds nothing to what is already conveyed by the value of p; in fact, this dichotomization of p-values makes matters worse.

For example, no p-value can reveal the plausibility, presence, truth, or importance of an association or effect. Therefore, a label of statistical signifficance does not mean or imply that an association or effect is highly probable, real, true, or important. Nor does a label of statistical nonsignifficance lead to the association or effect being improbable, absent, false, or unimportant. Yet the dichotomization into "signifficant" and "not signifficant" is taken as an imprimatur of authority on these characteristics. In a world without bright lines, on the other hand, it becomes untenable to assert dramatic differences in interpretation from inconsequential differences in estimates. As Gelman and Stern (2006) famously observed, the difference between "signifficant" and "not signifficant" is not itself statistically signifficant.

Furthermore, this false split into "worthy" and "unworthy" results leads to the selective reporting and publishing of results based on their statistical signifficance—the so-called "ffle drawer problem" (Rosenthal 1979). And the dichotomized reporting problem extends beyond just publication, notes Amrhein, Traffmow, and Greenland (2019): when authors use p-value thresholds to select which ffndings to discuss in their papers, "their conclusions and what is reported in subsequent news and reviews will be biased…Such selective attention based on study outcomes will therefore not only distort the literature but will slant published descriptions of study results—biasing the summary descriptions reported to practicing professionals and the general public." For the integrity of scientifc publishing and research dissemination, therefore, whether a p-value passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight.

To be clear, the problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups, based on arbitrary p-value thresholds. Similarly, we need to stop using conffdence intervals as another means of dichotomizing (based, on whether a null value falls within the interval). And, to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures (such as Bayes factors).

Despite the limitations of p-values (as noted in Principles 5 and 6 of the ASA statement), however, we are not recommending that the calculation and use of continuous p-values be discontinued. Where p-values are used, they should be reported as continuous quantities (e.g., $p = 0.08$). They should also be described in language stating what the value means in the scientifc context. We believe that a reasonable prerequisite for reporting any p-value is the ability to interpret it appropriately. We say more about this in Section 3.3.

To move forward to a world beyond "$p < 0.05$," we must recognize afresh that statistical inference is not—and never has been—equivalent to scientifc inference (Hubbard, Haig, and Parsa 2019; Ziliak 2019). However, looking to statistical signifficance for a marker of scientifc observations' credibility has created a guise of equivalency. Moving beyond "statistical signifficance" opens researchers to the real signifficance of statistics, which is "the science of learning from data, and of measuring, controlling, and communicating uncertainty" (Davidian and Louis 2012).

In sum, "statistically signifficant"—don't say it and don't use it.

3. There Are Many Do's

With the don'ts out of the way, we can ffnally discuss ideas for specifc, positive, constructive actions. We have a massive list of them in the seventh section of this editorialfffln that section, the authors of all the articles in this special issue each provide their own short set of do's. Those lists, and the rest of this editorial, will help you navigate the substantial collection of articles that follows.

Because of the size of this collection, we take the liberty here of distilling our readings of the articles into a summary of what can be done to move beyond "$p < 0.05$." You will ffnd the rich details in the articles themselves.

What you will NOT find in this issue is one solution that majestically replaces the outsized role that statistical significance has come to play. The statistical community has not yet converged on a simple paradigm for the use of statistical inference in scientifc research—and in fact it may never do so. A one-size-ffts-all approach to statistical inference is an inappropriate expectation, even affer the dust settles from our current remodeling of statistical practice (Tong 2019). Yet solid principles for the use of statistics do exist, and they are well explained in this special issue.

We summarize our recommendations in two sentences totaling seven words: "**A**ccept uncertainty. Be **t**houghtful, **o**pen, and **m**odest." Remember "ATOM."

3.1. Accept Uncertainty

Uncertainty exists everywhere in research. And, just like with the frigid weather in a Wisconsin winter, there are those who will ffee from it, trying to hide in warmer havens elsewhere. Others, however, accept and even delight in the omnipresent cold; these are the ones who buy the right gear and bravely take full advantage of all the wonders of a challenging climate. Signiffcance tests and dichotomized *p*-values have turned many researchers into scientiffc snowbirds, trying to avoid dealing with uncertainty by escaping to a "happy place" where results are either statistically signiffcant or not. In the real world, data provide a noisy signal. Variation, one of the causes of uncertainty, is everywhere. Exact replication is diff cult to achieve. So it is time to get the right (statistical) gear and "move toward a greater acceptance of uncertainty and embracing of variation" (Gelman 2016).

Statistical methods do not rid data of their uncertainty. "Statistics," Gelman (2016) says, "is offen sold as a sort of alchemy that transmutes randomness into certainty, an 'uncertainty laundering' that begins with data and concludes with success as measured by statistical signiffcance." To accept uncertainty requires that we "treat statistical results as being much more incomplete and uncertain than is currently the norm" (Amrhein, Traffmow, and Greenland 2019). We must "countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error" (Calin-Jageman and Cumming 2019).

"Accept uncertainty and embrace variation in effects," advise McShane et al. in Section 7 of this editorial. "[W]e can learn much (indeed, more) about the world by forsaking the false promise of certainty offered by dichotomous declarations of truth or falsity—binary statements about there being 'an effect' or 'no effect'—based on some *p*-value or other statistical threshold being attained."

We can make acceptance of uncertainty more natural to our thinking by accompanying every point estimate in our research with a measure of its uncertainty such as a standard error or interval estimate. Reporting and interpreting point and interval estimates should be routine. However, simplistic use of conffdence intervals as a measurement of uncertainty leads to the same bad outcomes as use of statistical signiffcance (especially, a focus on whether such intervals include or exclude the "null hypothesis value"). Instead, Greenland (2019) and Amrhein, Traffmow, and Greenland (2019) encourage thinking of conffdence intervals as "compatibility intervals," which use *p*-values to show the effect sizes that are most compatible with the data under the given model.

How will **accepting uncertainty** change anything? To begin, it will prompt us to seek better measures, more sensitive designs, and larger samples, all of which increase the rigor of research. It also helps us **be modest** (the fourth of our four principles, on which we will expand in Section 3.4) and encourages "meta-analytic thinking" (Cumming 2014). Accepting uncertainty as inevitable is a natural antidote to the seductive certainty falsely promised by statistical signiffcance. With this new outlook, we will naturally seek out replications and the integration of evidence through meta-analyses, which usually requires point and interval estimates from contributing studies. This will in turn give us more precise overall estimates for our effects and associations. And this is what will lead to the best research-based guidance for practical decisions.

Accepting uncertainty leads us to **be thoughtful**, the second of our four principles.

3.2. Be Thoughtful

What do we mean by this exhortation to "be thoughtful"? Researchers already clearly put much thought into their work. We are not accusing anyone of laziness. Rather, we are envisioning a sort of "statistical thoughtfulness." In this perspective, statistically **thoughtful researchers** begin above all else with clearly expressed objectives. They recognize when they are doing exploratory studies and when they are doing more rigidly pre-planned studies. They invest in producing solid data. They consider not one but a multitude of data analysis techniques. And they think about so much more.

3.2.1. Thoughtfulness in the Big Picture

"[M]ost scientiffc research is exploratory in nature," Tong (2019) contends. "[T]he design, conduct, and analysis of a study are necessarily ffexible, and must be open to the discovery of unexpected patterns that prompt new questions and hypotheses. In this context, statistical modeling can be exceedingly useful for elucidating patterns in the data, and researcher degrees of freedom can be helpful and even essential, though they still carry the risk of overffitting. The price of allowing this ffexibility is that the validity of any resulting statistical inferences is undermined."

Calin-Jageman and Cumming (2019) caution that "in practice the dividing line between planned and exploratory research can be diff cult to maintain. Indeed, exploratory ffndings have a slippery way of 'transforming' into planned ffndings as the research process progresses." At the bottom of that slippery slope one offen ffnds results that don't reproduce.

Anderson (2019) proposes three questions **thoughtful researchers** asked thoughtful researchers evaluating research results: What are the practical implications of the estimate? How precise is the estimate? And is the model correctly speciffed? The latter question leads naturally to three more: Are the modeling assumptions understood? Are these assumptions valid? And do the key results hold up when other modeling choices are made? Anderson further notes, "Modeling assumptions (including all the choices from model speciffcation to sample selection and the handling of data issues) should be suff ciently documented so independent parties can critique, and replicate, the work."

Drawing on archival research done at the Guinness Archives in Dublin, Ziliak (2019) emerges with ten "*G*-values" he believes we all wish to maximize in research. That is, we want large *G*-values, not small *p*-values. The ten principles of Ziliak's "Guinnessometrics" are derived primarily from his examination of experiments conducted by statistician William Sealy Gosset while working as Head Brewer for Guinness. Gosset took an economic approach to the logic of uncertainty, preferring balanced designs over random ones and estimation of gambles over bright-line "testing." Take, for example, Ziliak's *G*-value 10: "Consider purpose of the inquiry, and compare with best

practice," in the spirit of what farmers and brewers must do. The purpose is generally NOT to falsify a null hypothesis, says Ziliak. Ask what is at stake, he advises, and determine what magnitudes of change are humanly or scientifically meaningful in context.

Pogrow (2019) offers an approach based on practical benefit rather than statistical or practical significance. This approach is especially useful, he says, for assessing whether interventions in complex organizations (such as hospitals and schools) are effective, and also for increasing the likelihood that the observed benefits will replicate in subsequent research and in clinical practice. In this approach, "practical benefit" recognizes that reliance on small effect sizes can be as problematic as relying on *p*-values.

Thoughtful research prioritizes sound data production by putting energy into the careful planning, design, and execution of the study (Tong 2019).

Locascio (2019) urges researchers to be prepared for a new publishing model that evaluates their research based on the importance of the questions being asked and the methods used to answer them, rather than the outcomes obtained.

3.2.2. Thoughtfulness Through Context and Prior Knowledge

Thoughtful research considers the scientific context and prior evidence. In this regard, a declaration of statistical significance is the antithesis of thoughtfulness: it says nothing about practical importance, and it ignores what previous studies have contributed to our knowledge.

Thoughtful research looks ahead to prospective outcomes in the context of theory and previous research. Researchers would do well to ask, *What do we already know, and how certain are we in what we know?* And building on that and on the field's theory, *what magnitudes of difference, odds ratios, or other effect sizes are practically important?* These questions would naturally lead a researcher, for example, to use existing evidence from a literature review to identify specifically the findings that would be practically important for the key outcomes under study.

Thoughtful research includes careful consideration of the definition of a meaningful effect size. As a researcher you should communicate this up front, before data are collected and analyzed. Afterwards is just too late; it is dangerously easy to justify observed results after the fact and to overinterpret trivial effect sizes as being meaningful. Many authors in this special issue argue that consideration of the effect size and its "scientific meaningfulness" is essential for reliable inference (e.g., Blume et al. 2019; Betensky 2019). This concern is also addressed in the literature on equivalence testing (Wellek 2017).

Thoughtful research considers "related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain…without giving priority to *p*-values or other purely statistical measures" (McShane et al. 2019).

Thoughtful researchers "use a toolbox of statistical techniques, employ good judgment, and keep an eye on developments in statistical and data science," conclude Heck and Krueger (2019), who demonstrate how the *p*-value can be useful to researchers as a heuristic.

3.2.3. Thoughtful Alternatives and Complements to P-Values

Thoughtful research considers multiple approaches for solving problems. This special issue includes some ideas for supplementing or replacing *p*-values. Here is a short summary of some of them, with a few technical details:

Amrhein, Trafimow, and Greenland (2019) and Greenland (2019) advise that null *p*-values should be supplemented with a *p*-value from a test of a pre-specified alternative (such as a minimal important effect size). To reduce confusion with posterior probabilities and better portray evidential value, they further advise that *p*-values be transformed into *s*-values (Shannon information, surprisal, or binary logworth) $s = -\log_2(p)$. This measure of evidence aff rms other arguments that the evidence against a hypothesis contained in the *p*-value is not nearly as strong as is believed by many researchers. The change of scale also moves users away from probability misinterpretations of the *p*-value.

Blume et al. (2019) offer a "second generation *p*-value (SGPV)," the characteristics of which mimic or improve upon those of *p*-values but take practical significance into account. The null hypothesis from which an SGPV is computed is a composite hypothesis representing a range of differences that would be practically or scientifically inconsequential, as in equivalence testing (Wellek 2017). This range is determined in advance by the experimenters. When the SGPV is 1, the data only support null hypotheses; when the SGPV is 0, the data are incompatible with any of the null hypotheses. SGPVs between 0 and 1 are inconclusive at varying levels (maximally inconclusive at or near SGPV = 0.5.) Blume et al. illustrate how the SGPV provides a straightforward and useful descriptive summary of the data. They argue that it eliminates the problem of how classical statistical significance does not imply scientific relevance, it lowers false discovery rates, and its conclusions are more likely to reproduce in subsequent studies.

The "analysis of credibility"(AnCred) is promoted by Matthews (2019). This approach takes account of both the width of the confidence interval and the location of its bounds when assessing weight of evidence. AnCred assesses the credibility of inferences based on the confidence interval by determining the level of prior evidence needed for a new finding to provide credible evidence for a nonzero effect. If this required level of prior evidence is supported by current knowledge and insight, Matthews calls the new result "credible evidence for a non-zero effect," irrespective of its statistical significance/nonsignificance.

Colquhoun (2019) proposes continuing the use of continuous *p*-values, but only in conjunction with the "false positive risk (FPR)." The FPR answers the question, "If you observe a 'significant' *p*-value after doing a single unbiased experiment, what is the probability that your result is a false positive?" It tells you what most people mistakenly still think the *p*-value does, Colquhoun says. The problem, however, is that to calculate the FPR you need to specify the prior probability that an effect is real, and it's rare to know this. Colquhoun suggests that the FPR could be calculated with a prior probability of 0.5, the largest value reasonable to assume in the absence of hard prior data. The FPR found this way is in a sense the minimum false positive risk (mFPR); less plausible hypotheses (prior probabilities below 0.5) would give even bigger FPRs, Colquhoun says, but the

mFPR would be a big improvement on reporting a p-value alone. He points out that p-values near 0.05 are, under a variety of assumptions, associated with minimum false positive risks of 20–30%, which should stop a researcher from making too big a claim about the "statistical signifficance" of such a result.

Benjamin and Berger (2019) propose a different supplement to the null p-value. The Bayes factor bound (BFB)—which under typically plausible assumptions is the value $1/(-ep \ln p)$—represents the upper bound of the ratio of data-based odds of the alternative hypothesis to the null hypothesis. Benjamin and Berger advise that the BFB should be reported along with the continuous p-value. This is an incomplete step toward revising practice, they argue, but one that at least confronts the researcher with the maximum possible odds that the alternative hypothesis is true—which is what researchers offen think they are getting with a p-value. The BFB, like the FPR, offen clariffes that the evidence against the null hypothesis contained in the p-value is not nearly as strong as is believed by many researchers.

Goodman, Spruill, and Komaroff (2019) propose a two-stage approach to inference, requiring both a small p-value below a pre-speciffed level and a pre-speciffed suff ciently large effect size before declaring a result "signiffcant." They argue that this method has improved performance relative to use of dichotomized p-values alone.

Gannon, Pereira, and Polpo (2019) have developed a testing procedure combining frequentist and Bayesian tools to provide a signiffcance level that is a function of sample size.

Manski (2019) and Manski and Tetenov (2019) urge a return to the use of statistical decision theory, which they say has largely been forgotten. Statistical decision theory is not based on p-value thresholds and readily distinguishes between statistical and clinical signiffcance.

Billheimer (2019) suggests abandoning inference about parameters, which are frequently hypothetical quantities used to idealize a problem. Instead, he proposes focusing on the prediction of future observables, and their associated uncertainty, as a means to improving science and decision-making.

3.2.4. Thoughtful Communication of Confidence

Be thoughtful and clear about the level of conffdence or credibility that is present in statistical results.

Amrhein, Traffmow, and Greenland (2019) and Greenland (2019) argue that the use of words like "signiffcance" in conjunction with p-values and "conffdence" with interval estimates misleads users into overconffdent claims. They propose that researchers think of p-values as measuring the compatibility between hypotheses and data, and interpret interval estimates as "compatibility intervals."

In what may be a controversial proposal, Goodman (2018) suggests requiring "that any researcher making a claim in a study accompany it with their estimate of the chance that the claim is true." Goodman calls this the conffdence index. For example, along with stating "This drug is associated with elevated risk of a heart attack, relative risk (RR) $= 2.4$, $p = 0.03$," Goodman says investigators might add a statement such as "There is an 80% chance that this drug raises the risk, and a 60% chance that the risk is at least doubled." Goodman acknowledges, "Although simple on paper, requiring a conffdence index would entail a profound overhaul of scientiffc and statistical practice."

In a similar vein, Hubbard and Carriquiry (2019) urge that researchers prominently display the probability the hypothesis is true or a probability distribution of an effect size, or provide suff cient information for future researchers and policy makers to compute it. The authors further describe why such a probability is necessary for decision making, how it could be estimated by using historical rates of reproduction of ffndings, and how this same process can be part of continuous "quality control" for science.

Being thoughtful in our approach to research will lead us to **be open** in our design, conduct, and presentation of it as well.

3.3. Be Open

We envision **openness** as embracing certain positive practices in the development and presentation of research work.

3.3.1. Openness to Transparency and to the Role of Expert Judgment

First, we repeat off-repeated advice: **Be open** to "open science" practices. Calin-Jageman and Cumming (2019), Locascio (2019), and others in this special issue urge adherence to practices such as public pre-registration of methods, transparency and completeness in reporting, shared data and code, and even pre-registered ("results-blind") review. Completeness in reporting, for example, requires not only describing all analyses performed but also presenting all ffndings obtained, without regard to statistical signiffcance or any such criterion.

Openness also includes understanding and accepting the role of expert judgment, which enters the practice of statistical inference and decision-making in numerous ways (O'Hagan 2019). "Indeed, there is essentially no aspect of scientiffc investigation in which judgment is not required," O'Hagan observes. "Judgment is necessarily subjective, but should be made as carefully, as objectively, and as scientiffcally as possible."

Subjectivity is involved in any statistical analysis, Bayesian or frequentist. Gelman and Hennig (2017) observe, "Personal decision making cannot be avoided in statistical data analysis and, for want of approaches to justify such decisions, the pursuit of objectivity degenerates easily to a pursuit to merely *appear* objective." One might say that subjectivity is not a problem; it is part of the solution.

Acknowledging this, Brownstein et al. (2019) point out that expert judgment and knowledge are required in all stages of the scientiffc method. They examine the roles of expert judgment throughout the scientiffc process, especially regarding the integration of statistical and content expertise. "All researchers, irrespective of their philosophy or practice, use expert judgment in developing models and interpreting results," say Brownstein et al. "We must accept that there is subjectivity in every stage of scientiffc inquiry, but objectivity is nevertheless the fundamental goal. Therefore, we should base judgments on evidence and careful reasoning, and seek wherever possible to eliminate potential sources of bias."

How does one rigorously elicit expert knowledge and judgment in an effective, unbiased, and transparent way? O'Hagan (2019) addresses this, discussing protocols to elicit expert knowledge in an unbiased and as scientifically sound was as possible. It is also important for such elicited knowledge to be examined critically, comparing it to actual study results being an important diagnostic step.

3.3.2. Openness in Communication

Be open in your reporting. Report *p*-values as continuous, descriptive statistics, as we explain in Section 2. We realize that this leaves researchers without their familiar bright line anchors. Yet if we were to propose a universal template for presenting and interpreting continuous *p*-values we would violate our own principlesffRather, we believe that the thoughtful use and interpretation of *p*-values will never adhere to a rigid rulebook, and will instead inevitably vary from study to study. Despite these caveats, we can offer recommendations for sound practices, as described below.

In all instances, regardless of the value taken by *p* or any other statistic, consider what McShane et al. (2019) call the "currently subordinate factors"—the factors that should no longer be subordinate to "$p < 0.05$." These include relevant prior evidence, plausibility of mechanism, study design and data quality, and the real-world costs and benefits that determine what effects are scientifically important. The scientific context of your study matters, they say, and this should guide your interpretation.

When using *p*-values, remember not only Principle 5 of the ASA statement: "A *p*-value…does not measure the size of an effect or the importance of a result" but also Principle 6: "By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis." Despite these limitations, if you present *p*-values, do so for more than one hypothesized value of your variable of interest (Fraser 2019; Greenland 2019), such as 0 and at least one plausible, relevant alternative, such as the minimum practically important effect size (which should be determined before analyzing the data).

Betensky (2019) also reminds us to interpret the *p*-value in the context of sample size and meaningful effect size.

Instead of *p*, you might consider presenting the *s*-value (Greenland 2019), which is described in Section 3.2. As noted in Section 3.1, you might present a confidence interval. Sound practices in the interpretation of confidence intervals include (1) discussing both the upper and lower limits and whether they have different practical implications, (2) paying no particular attention to whether the interval includes the null value, and (3) remembering that an interval is itself an estimate subject to error and generally provides only a rough indication of uncertainty given that all of the assumptions used to create it are correct and, thus, for example, does not "rule out" values outside the interval. Amrhein, Trafimow, and Greenland (2019) suggest that interval estimates be interpreted as "compatibility" intervals rather than as "confidence" intervals, showing the values that are most compatible with the data, under the model used to compute the interval. They argue that such an interpretation and the practices outlined here can help guard against overconfidence.

It is worth noting that Tong (2019) disagrees with using *p*-values as descriptive statistics. "Divorced from the probability claims attached to such quantities (confidence levels, nominal Type I errors, and so on), there is no longer any reason to privilege such quantities over descriptive statistics that more directly characterize the data at hand." He further states, "Methods with alleged generality, such as the *p*-value or Bayes factor, should be avoided in favor of discipline- and problem-specific solutions that can be designed to be fit for purpose."

Failing to **be open** in reporting leads to publication bias. Ioannidis (2019) notes the high level of selection bias prevalent in biomedical journals. He defines "selection" as "the collection of choices that lead from the planning of a study to the reporting of *p*-values." As an illustration of one form of selection bias, Ioannidis compared "the set of *p*-values reported in the full text of an article with the set of *p*-values reported in the abstract." The main finding, he says, "was that *p*-values chosen for the abstract tended to show greater significance than those reported in the text, and that the gradient was more pronounced in some types of journals and types of designs." Ioannidis notes, however, that selection bias "can be present regardless of the approach to inference used." He argues that in the long run, "the only direct protection must come from standards for reproducible research."

To **be open**, remember that one study is rarely enough. The words "a groundbreaking new study" might be loved by news writers but must be resisted by researchers. Breaking ground is only the first step in building a house. It will be suitable for habitation only after much more hard work.

Be open by providing sufficient information so that other researchers can execute meaningful alternative analyses. van Dongen et al. (2019) provide an illustrative example of such alternative analyses by different groups attacking the same problem.

Being open goes hand in hand with **being modest.**

3.4. Be Modest

Researchers of any ilk may rarely advertise their personal modesty. Yet the most successful ones cultivate a practice of **being modest** throughout their research, by understanding and clearly expressing the limitations of their work.

Being modest requires a reality check (Amrhein, Trafimow, and Greenland 2019). "A core problem," they observe, "is that both scientists and the public confound statistics with reality. But statistical inference is a thought experiment, describing the predictive performance of models about reality. Of necessity, these models are extremely simplified relative to the complexities of actual study conduct and of the reality being studied. Statistical results must eventually mislead us when they are used and communicated as if they present this complex reality, rather than a model for it. This is not a problem of our statistical methods. It is a problem of interpretation and communication of results."

Be modest in recognizing there is not a "true statistical model" underlying every problem, which is why it is wise to **thoughtfully** consider many possible models (Lavine 2019). Rougier (2019) calls on researchers to "recognize that behind every choice of null distribution and test statistic, there lurks

a plausible family of alternative hypotheses, which can provide more insight into the null distribution."

p-values, conffdence intervals, and other statistical measures are all uncertain. Treating them otherwise is immodest overconffdence.

Remember that statistical tools have their limitations. Rose and McGuire (2019) show how use of stepwise regression in health care settings can lead to policies that are unfair.

Remember also that the amount of evidence for or against a hypothesis provided by *p*-values near the ubiquitous $p < 0.05$ threshold (Johnson 2019) is usually much less than you think (Benjamin and Berger 2019; Colquhoun 2019; Greenland 2019).

Be modest about the role of statistical inference in scientiffc inference. "Scientiffc inference is a far broader concept than statistical inference," says Hubbard, Haig, and Parsa (2019). "A major focus of scientiffc inference can be viewed as the pursuit of *significant sameness,* meaning replicable and empirically generalizable results among phenomena. Regrettably, the obsession with users of statistical inference to report *significant difierences* in data sets actively thwarts cumulative knowledge development."

The nexus of **openness** and **modesty** is to report everything while at the same time not concluding anything from a single study with unwarranted certainty. Because of the strong desire to inform and be informed, there is a relentless demand to state results with certainty. Again, **accept uncertainty** and embrace variation in associations and effects, because they are always there, like it or not. Understand that expressions of uncertainty are themselves uncertain. Accept that one study is rarely deffnitive, so encourage, sponsor, conduct, and publish replication studies. Then, use meta-analysis, evidence reviews, and Bayesian methods to synthesize evidence across studies.

Resist the urge to overreach in the generalizability of claims, Watch out for pressure to embellish the abstract or the press release. If the study's limitations are expressed in the paper but not in the abstract, they may never be read.

Be modest by encouraging others to reproduce your work. Of course, for it to be reproduced readily, you will necessarily have been **thoughtful** in conducting the research and **open** in presenting it.

Hubbard and Carriquiry (see their "do list" in Section 7) suggest encouraging reproduction of research by giving "a byline status for researchers who reproduce studies." They would like to see digital versions of papers dynamically updated to display "Reproduced by…." below original research authors' names or "not yet reproduced" until it is reproduced.

Indeed, when it comes to reproducibility, Amrhein, Traffmow, and Greenland (2019) demand that we **be modest** in our expectations. "An important role for statistics in research is the summary and accumulation of information," they say. "If replications do not ffnd the same results, this is not necessarily a crisis, but is part of a natural process by which science evolves. The goal of scientiffc methodology should be to direct this evolution toward ever more accurate descriptions of the world and how it works, not toward ever more publication of inferences, conclusions, or decisions."

Referring to replication studies in psychology, McShane et al. (2019) recommend that future large-scale replication projects "should follow the 'one phenomenon, many studies' approach of the Many Labs project and Registered Replication Reports rather than the 'many phenomena, one study' approach of the Open Science Collaboration project. In doing so, they should systematically vary method factors across the laboratories involved in the project." This approach helps achieve the goals of Amrhein, Traffmow, and Greenland (2019) by increasing understanding of why and when results replicate or fail to do so, yielding more accurate descriptions of the world and how it works. It also speaks to signiffcant sameness versus signiffcant difference a la Hubbard, Haig, and Parsa (2019).

Kennedy-Shaffer's (2019) historical perspective on statistical signiffcance reminds us to **be modest**, by prompting us to recall how the current state of affairs in *p*-values has come to be.

Finally, **be modest** by recognizing that different readers may have very different stakes on the results of your analysis, which means you should try to take the role of a neutral judge rather than an advocate for any hypothesis. This can be done, for example, by pairing every null *p*-value with a *p*-value testing an equally reasonable alternative, and by discussing the endpoints of every interval estimate (not only whether it contains the null).

Accept that both scientiffc inference and statistical inference are hard, and understand that no knowledge will be eff ciently advanced using simplistic, mechanical rules and procedures. Accept also that pure objectivity is an unattainable goal—no matter how laudable—and that both subjectivity and expert judgment are intrinsic to the conduct of science and statistics. Accept that there will always be uncertainty, and be **t**houghtful, **o**pen, and **m**odest. ATOM.

And to push this acronym further, we argue in the next section that **i**nstitutional **c**hange is needed, so we put forward that change is needed at the ATOMIC level. Let's go.

4. Editorial, Educational and Other Institutional Practices Will Have to Change

Institutional reform is necessary for moving beyond statistical signiffcance in any context—whether journals, education, academic incentive systems, or others. Several papers in this special issue focus on reform.

Goodman (2019) notes considerable social change is needed in academic institutions, in journals, and among funding and regulatory agencies. He suggests (see Section 7) partnering "with science reform movements and reformers within disciplines, journals, funding agencies and regulators to promote and reward 'reproducible' science and diminish the impact of statistical signiffcance on publication, funding and promotion." Similarly, Colquhoun (2019) says, "In the end, the only way to solve the problem of reproducibility is to do more replication and to reduce the incentives that are imposed on scientists to produce unreliable work. The publish-or-perish culture has damaged science, as has the judgment of their work by silly metrics."

Traffmow (2019), who added energy to the discussion of *p*-values a few years ago by banning them from the journal he edits (Fricker et al. 2019), suggests ffve "nonobvious changes" to editorial practice. These suggestions, which demand reevaluating traditional practices in editorial policy, will not be trivial to implement but would result in massive change in some journals.

Locascio (2017, 2019) suggests that evaluation of manuscripts for publication should be "results-blind." That is, manuscripts should be assessed for suitability for publication based on the substantive importance of the research without regard to their reported results. Kmetz (2019) supports this approach as well and says that it would be a huge benefft for reviewers, "freeing [them] from their offen thankless present jobs and instead allowing them to review research designs for their potential to provide useful knowledge." (See also "registered reports" from the Center for Open Science (*https://cos.io/rr/?_ga=2.184185454.979594832.1547755516-1193527346.1457026171*) and "registered replication reports" from the Association for Psychological Science (*https://www.psychologicalscience.org/publications/replication*) in relation to this concept.)

Amrhein, Traffmow, and Greenland (2019) ask if results-blind publishing means that anything goes, and then answer aff rmatively: "Everything should be published in some form if whatever we measured made sense *before we obtained the data* because it was connected in a potentially useful way to some research question." Journal editors, they say, "should be proud about [their] exhaustive methods sections" and base their decisions about the suitability of a study for publication "on the quality of its materials and methods rather than on results and conclusions; the quality of the presentation of the latter is only judged affer it is determined that the study is valuable based on its materials and methods."

A "variation on this theme is *pre-registered replication*, where a *replication* study, rather than the original study, is subject to strict pre-registration (e.g., Gelman 2015)," says Tong (2019). "A broader vision of this idea (Mogil and Macleod 2017) is to carry out a whole series of exploratory experiments *without* any formal statistical inference, and summarize the results by descriptive statistics (including graphics) or even just disclosure of the raw data. When results from this series of experiments converge to a single working hypothesis, it can *then* be subjected to a pre-registered, randomized and blinded, appropriately powered conffrmatory experiment, carried out by another laboratory, in which valid statistical inference may be made."

Hurlbert, Levine, and Utts (2019) urge abandoning the use of "statistically signiffcant" in all its forms and encourage journals to provide instructions to authors along these lines: "There is now wide agreement among many statisticians who have studied the issue that for reporting of statistical tests yielding p-values it is illogical and inappropriate to dichotomize the p-scale and describe results as 'signiffcant' and 'nonsigniffcant.' Authors are strongly discouraged from continuing this never justiffed practice that originated from confusions in the early history of modern statistics."

Hurlbert, Levine, and Utts (2019) also urge that the *ASA Statement on P-Values and Statistical Significance* "be sent to the editor-in-chief of every journal in the natural, behavioral and social sciences for forwarding to their respective editorial boards and stables of manuscript reviewers. That would be a good way to quickly improve statistical understanding and practice." Kmetz (2019) suggests referring to the ASA statement whenever submitting a paper or revision to any editor, peer reviewer, or prospective reader. Hurlbert et al. encourage a "community grassroots effort" to encourage change in journal procedures.

Campbell and Gustafson (2019) propose a statistical model for evaluating publication policies in terms of weighing novelty of studies (and the likelihood of those studies subsequently being found false) against pre-speciffed study power. They observe that "no publication policy will be perfect. Science is inherently challenging and we must always be willing to accept that a certain proportion of research is potentially false."

Statistics education will require major changes at all levels to move to a post "$p < 0.05$" world. Two papers in this special issue make a speciffc start in that direction (Maurer et al. 2019; Steel, Liermann, and Guttorp 2019), but we hope that volumes will be written on this topic in other venues. We are excited that, with support from the ASA, the US Conference on Teaching Statistics (USCOTS) will focus its 2019 meeting on teaching inference.

The change that needs to happen demands change to editorial practice, to the teaching of statistics at every level where inference is taught, and to much more. However…

5. It Is Going to Take Work, and It Is Going to Take Time

If it were easy, it would have already been done, because as we have noted, this is nowhere near the ffrst time the alarm has been sounded.

Why is eliminating the use of p-values as a truth arbiter so hard? "The basic explanation is neither philosophical nor scientiffc, but sociologic; everyone uses them," says Goodman (2019). "It's the same reason we can use money. When everyone believes in something's value, we can use it for real things; money for food, and p-values for knowledge claims, publication, funding, and promotion. It doesn't matter if the p-value doesn't mean what people think it means; it becomes valuable because of what it buys."

Goodman observes that statisticians alone cannot address the problem, and that "any approach involving only statisticians will not succeed." He calls on statisticians to ally themselves "both with scientists in other ffelds and with broader based, multidisciplinary scientiffc reform movements. What statisticians can do within our own discipline is important, but to effectively disseminate or implement virtually any method or policy, we need partners."

"The loci of inffuence," Goodman says, "include journals, scientiffc lay and professional media (including social media), research funders, healthcare payors, technology assessors, regulators, academic institutions, the private sector, and professional societies. They also can include policy or informational entities like the National Academies…as well as various other science advisory bodies across the government. Increasingly, they are also including non-traditional science reform organizations comprised both of scientists and of the science literate lay public…and a broad base of health or science advocacy groups…"

It is no wonder, then, that the problem has persisted for so long. And persist it hasff Hubbard (2019) looked at citation-count data on twenty-ffve articles and books severely critical of the effect of null hypothesis signiffcance testing (NHST) on good science. Though issues were well known, Hubbard says, this did nothing to stem NHST usage over time.

Greenland (personal communication, January 25, 2019) notes that cognitive biases and perverse incentives to offer ffrm conclusions where none are warranted can warp the use of any method. "The core human and systemic problems are not addressed by shiffing blame to *p*-values and pushing alternatives as magic cures—especially alternatives that have been subject to little or no comparative evaluation in either classrooms or practice," Greenland said. "What we need now is to move beyond debating only our methods and their interpretations, to concrete proposals for elimination of systemic problems such as pressure to produce noteworthy ffndings rather than to produce reliable studies and analyses. Review and provisional acceptance of reports before their results are given to the journal (Locascio 2019) is one way to address that pressure, but more ideas are needed since review of promotions and funding applications cannot be so blinded. The challenges of how to deal with human biases and incentives may be the most diff cult we must face." Supporting this view is McShane and Gal's (2016, 2017) empirical demonstration of cognitive dichotomization errors among biomedical and social science researchers—and even among statisticians.

Challenges for editors and reviewers are many. Here's an example: Fricker et al. (2019) observed that when *p*-values were suspended from the journal *Basic and Applied Social Psychology* authors tended to overstate conclusions.

With all the challenges, how do we get from here to there, from a "$p < 0.05$" world to a post "$p < 0.05$" world?

Matthews (2019) notes that "Any proposal encouraging changes in inferential practice must accept the ubiquity of NHST.…Pragmatism suggests, therefore, that the best hope of achieving a change in practice lies in offering inferential tools that can be used alongside the concepts of NHST, adding value to them while mitigating their most egregious features."

Benjamin and Berger (2019) propose three practices to help researchers during the transition away from use of statistical signiffcance. "…[O]ur goal is to suggest minimal changes that would require little effort for the scientiffc community to implement," they say. "Motivating this goal are our hope that easy (but impactful) changes might be adopted and our worry that more complicated changes could be resisted simply because they are perceived to be too diff cult for routine implementation."

Yet there is also concern that progress will stop affer a small step or two. Even some proponents of small steps are clear that those small steps still carry us far short of the destination.

For example, Matthews (2019) says that his proposed methodology "is not a panacea for the inferential ills of the research community." But that doesn't make it useless. It may "encourage researchers to move beyond NHST and explore the statistical armamentarium now available to answer the central question of research: what does our study tell us?" he says. It "provides a bridge between the dominant but ffawed NHST paradigm and the less familiar but more informative methods of Bayesian estimation."

Likewise, Benjamin and Berger (2019) observe, "In research communities that are deeply attached to reliance on '$p < 0.05$,' our recommendations will serve as initial steps away from this attachment. We emphasize that our recommendations are intended merely as initial, temporary steps and that many further steps will need to be taken to reach the ultimate destination: a holistic interpretation of statistical evidence that fully conforms to the principles laid out in the ASA Statement…"

Yet, like the authors of this editorial, not all authors in this special issue support gradual approaches with transitional methods.

Some (e.g., Amrhein, Traffmow, and Greenland 2019; Hurlbert, Levine, and Utts 2019; McShane et al. 2019) prefer to rip off the bandage and abandon use of statistical signiffcance altogether. In short, no more dichotomizing *p*-values into categories of "signiffcance." Notably, these authors do not suggest banning the use of *p*-values, but rather suggest using them descriptively, treating them as continuous, and assessing their weight or import with nuanced thinking, clear language, and full understanding of their properties.

So even when there is agreement on the destination, there is disagreement about what road to take. The questions around reform need consideration and debate. It might turn out that different ffelds take different roads.

The catalyst for change may well come from those people who fund, use, or depend on scientiffc research, say Calin-Jageman and Cumming (2019). They believe this change has not yet happened to the desired level because of "the cognitive opacity of the NHST approach: the counter-intuitive *p*-value (it's good when it is small), the mysterious null hypothesis (you want it to be false), and the eminently confusable Type I and Type II errors."

Reviewers of this editorial asked, as some readers of it will, is a *p*-value threshold ever okay to use? We asked some of the authors of articles in the special issue that question as well. Authors identiffed four general instances. Some allowed that, while *p*-value thresholds should not be used for inference, they might still be useful for applications such as industrial quality control, in which a highly automated decision rule is needed and the costs of erroneous decisions can be carefully weighed when specifying the threshold. Other authors suggested that such dichotomized use of *p*-values was acceptable in model-fftting and variable selection strategies, again as automated tools, this time for sorting through large numbers of potential models or variables. Still others pointed out that *p*-values with very low thresholds are used in ffelds such as physics, genomics, and imaging as a fflter for massive numbers of tests. The fourth instance can be described as "conffrmatory setting[s] where the study design and statistical analysis plan are speciffed prior to data collection, and then adhered to during and affer it" (Tong 2019). Tong argues these are the only proper settings for formal statistical inference. And Wellek (2017) says at present it is essential in these settings. "[B]inary decision making is indispensable in medicine and related ffelds," he says. "[A] radical rejection of the classical principles of statistical inference…is of virtually no help as long as no conclusively substantiated alternative can be offered."

Eliminating the declaration of "statistical signiffcance" based on $p < 0.05$ or other arbitrary thresholds will be easier in some venues than others. Most journals, if they are willing, could fairly rapidly implement editorial policies to effect these changes. Suggestions for how to do that are in this special issue of *The American Statistician*. However, regulatory agencies might require longer timelines for making changes. The U.S. Food and

Drug Administration (FDA), for example, has long established drug review procedures that involve comparing p-values to signifficance thresholds for Phase III drug trials. Many factors demand consideration, not the least of which is how to avoid turning every drug decision into a court battle. Goodman (2019) cautions that, even as we seek change, "we must respect the reason why the statistical procedures are there in the ffrst place." Perhaps the ASA could convene a panel of experts, internal and external to FDA, to provide a workable new paradigm. (See Ruberg et al. 2019, who argue for a Bayesian approach that employs data from other trials as a "prior" for Phase 3 trials.)

Change is needed. Change has been needed for decades. Change has been called for by others for quite a while. So…

6. Why Will Change Finally Happen Now?

In 1991, a conffuence of weather events created a monster storm that came to be known as "the perfect storm," entering popular culture through a book (Junger 1997) and a 2000 movie starring George Clooney. Concerns about reproducible science, falling public conffdence in science, and the initial impact of the ASA statement in heightening awareness of long-known problems created a perfect storm, in this case, a good storm of motivation to make lasting change. Indeed, such change was the intent of the ASA statement, and we expect this special issue of TAS will inject enough additional energy to the storm to make its impact widely felt.

We are not alone in this view. "60+ years of incisive criticism has not yet dethroned NHST as the dominant approach to inference in many ffelds of science," note Calin-Jageman and Cumming (2019). "Momentum, though, seems to ffnally be on the side of reform."

Goodman (2019) agrees: "The initial slow speed of progress should not be discouraging; that is how all broad-based social movements move forward and we should be playing the long game. But the ball is rolling downhill, the current generation is inspired and impatient to carry this forward."

So, let's do it. Let's move beyond "statistically signifficant," even if upheaval and disruption are inevitable for the time being. It's worth it. In a world beyond "$p < 0.05$," by breaking free from the bonds of statistical signifficance, statistics in science and policy will become more signifficant than ever.

7. Authors' Suggestions

The editors of this special TAS issue on statistical inference asked all the contact authors to help us summarize the guidance they provided in their papers by providing us a short list of do's. We asked them to be speciffc but concise and to be active—start each with a verb. Here is the complete list of the authors' responses, ordered as the papers appear in this special issue.

7.1. Getting to a Post "p < 0.05" Era

Ioannidis, J., What Have We (Not) Learnt From Millions of Scientific Papers With p-Values?

1. Do not use p-values, unless you have clearly thought about the need to use them and they still seem the best choice.
2. Do not favor "statistically signifficant" results.
3. Do be highly skeptical about "statistically signifficant" results at the 0.05 level.

Goodman, S., Why Is Getting Rid of p-Values So Hard? Musings on Science and Statistics

1. Partner with science reform movements and reformers within disciplines, journals, funding agencies and regulators to promote and reward reproducible science and diminish the impact of statistical signifficance on publication, funding and promotion.
2. Speak to and write for the multifarious array of scientiffc disciplines, showing how statistical uncertainty and reasoning can be conveyed in non-"bright-line" ways both with conventional and alternative approaches. This should be done not just in didactic articles, but also in original or reanalyzed research, to demonstrate that it is publishable.
3. Promote, teach and conduct meta-research within many individual scientiffc disciplines to demonstrate the adverse effects in each of over-reliance on and misinterpretation of p-values and signifficance verdicts in individual studies and the benefffts of emphasizing estimation and cumulative evidence.
4. Require reporting a quantitative measure of certainty—a "conffdence index"—that an observed relationship, or claim, is true. Change analysis goal from achieving signifficance to appropriately estimating this conffdence.
5. Develop and share teaching materials, soffware, and published case examples to help with all of the do's above, and to spread progress in one discipline to others.

Hubbard, R., Will the ASA's Efiorts to Improve Statistical Practice be Successful? Some Evidence to the Contrary

This list applies to the ASA and to the professional statistics community more generally.

1. Specify, where/if possible, those situations in which the p-value plays a clearly valuable role in data analysis and interpretation.
2. Contemplate issuing a statement abandoning the use of p-values in null hypothesis signifficance testing.

Kmetz, J., Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of p-Values

1. Refer to the ASA statement on p-values whenever submitting a paper or revision to any editor, peer reviewer, or prospective reader. Many in the ffeld do not know of this statement, and having the support of a prestigious organization when authoring any research document will help stop corrupt research from becoming even more dominant than it is.
2. Train graduate students and future researchers by having them reanalyze published studies and post their ffndings to appropriate websites or weblogs. This practice will benefft not only the students, but will benefft the professions, by increasing the amount of replicated (or nonreplicated) research available and readily accessible, and as well as reformer organizations that support replication.
3. Join one or more of the reformer organizations formed or forming in many research ffelds, and support and publicize their efforts to improve the quality of research practices.

4. Challenge editors and reviewers when they assert that incorrect practices and interpretations of research, consistent with existing null hypothesis signiffcance testing and beliefs regarding *p*-values, should be followed in papers submitted to their journals. Point out that new submissions have been prepared to be consistent with the ASA statement on *p*-values.
5. Promote emphasis on research quality rather than research quantity in universities and other institutions where professional advancement depends heavily on research "productivity," by following the practices recommended in this special journal edition. This recommendation will fall most heavily on those who have already achieved success in their ffelds, perhaps by following an approach quite different from that which led to their success; whatever the merits of that approach may have been, one objectionable outcome of it has been the production of voluminous corrupt research and creation of an environment that promotes and protects it. We must do better.

Hubbard, D., and Carriquiry, A., Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness and Relevance

1. Compute and prominently display the probability the hypothesis is true (or a probability distribution of an effect size) or provide suff cient information for future researchers and policy makers to compute it.
2. Promote publicly displayed quality control metrics within your ffeld—in particular, support tracking of reproduction studies and computing the "level 1" and even "level 2" priors as required for #1 above.
3. Promote a byline status for researchers who reproduce studies: Digital versions are dynamically updated to display "Reproduced by…." below original research authors' names or "Not yet reproduced" until it is reproduced.

Brownstein, N., Louis, T., O'Hagan, A., and Pendergast, J., The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making

1. Staff the study team with members who have the necessary knowledge, skills and experience—statistically, scientiffcally, and otherwise.
2. Include key members of the research team, including statisticians, in all scientiffc and administrative meetings.
3. Understand that subjective judgments are needed in all stages of a study.
4. Make all judgments as carefully and rigorously as possible and document each decision and rationale for transparency and reproducibility.
5. Use protocol-guided elicitation of judgments.
6. Statisticians speciffcally should:

   - Reffne oral and written communication skills.
   - Understand their multiple roles and obligations as collaborators.
   - Take an active leadership role as a member of the scientiffc team; contribute throughout all phases of the study.
   - Co-own the subject matter—understand a suff cient amount about the relevant science/policy to meld statistical and subject-area expertise.
   - Promote the expectation that your collaborators co-own statistical issues.
   - Write a statistical analysis plan for all analyses and track any changes to that plan over time.
   - Promote co-responsibility for data quality, security, and documentation.
   - Reduce unplanned and uncontrolled modeling/testing (HARK-ing, *p*-hacking); document all analyses.

O'Hagan, A., Expert Knowledge Elicitation: Subjective but Scientific

1. Elicit expert knowledge when data relating to a parameter of interest is weak, ambiguous or indirect.
2. Use a well-designed protocol, such as SHELF, to ensure expert knowledge is elicited in as scientiffc and unbiased a way as possible.

Kennedy-Shafier, L., Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing

1. Ensure that inference methods match intuitive understandings of statistical reasoning.
2. Reduce the computational burden for nonstatisticians using statistical methods.
3. Consider changing conditions of statistical and scientiffc inference in developing statistical methods.
4. Address uncertainty quantitatively and in ways that reward increased precision.

Hubbard, R., Haig, B. D., and Parsa, R. A., The Limited Role of Formal Statistical Inference in Scientific Inference

1. Teach readers that although deemed equivalent in the social, management, and biomedical sciences, formal methods of statistical inference and scientiffc inference are very different animals.
2. Show these readers that formal methods of statistical inference play only a restricted role in scientiffc inference.
3. Instruct researchers to pursue signiffcant *sameness* (i.e., replicable and empirically generalizable results) rather than signiffcant *difierences* in results.
4. Demonstrate how the pursuit of signiffcant differences actively impedes cumulative knowledge development.

McShane, B., Tackett, J., Böckenholt, U., and Gelman, A., Large Scale Replication Projects in Contemporary Psychological Research

1. When planning a replication study of a given psychological phenomenon, bear in mind that replication is complicated in psychological research because studies can never be direct or exact replications of one another, and thus heterogeneity—effect sizes that vary from one study of the phenomenon to the next—cannot be avoided.
2. Future large scale replication projects should follow the "one phenomenon, many studies" approach of the Many Labs project and Registered Replication Reports rather than the

"many phenomena, one study" approach of the Open Science Collaboration project. In doing so, they should systematically vary method factors across the laboratories involved in the project.
3. Researchers analyzing the data resulting from large scale replication projects should do so via a hierarchical (or multilevel) model fft to the totality of the individual-level observations. In doing so, all theoretical moderators should be modeled via covariates while all other potential moderators—that is, method factors—should induce variation (i.e., heterogeneity).
4. Assessments of replicability should not depend solely on estimates of effects, or worse, signiffcance tests based on them. Heterogeneity must also be an important consideration in assessing replicability.

7.2. Interpreting and Using p

Greenland, S., Valid p-Values Behave Exactly as They Should: Some Misleading Criticisms of p-Values and Their Resolution With s-Values

1. Replace any statements about statistical signiffcance of a result with the p-value from the test, and present the p-value as an equality, not an inequality. For example, if $p = 0.03$ then "…was statistically signiffcant" would be replaced by "…had $p = 0.03$," and "$p < 0.05$" would be replaced by "$p = 0.03$." (An exception: If p is so small that the accuracy becomes very poor then an inequality reffecting that limit is appropriate; e.g., depending on the sample size, p-values from normal or χ^2 approximations to discrete data offen lack even 1-digit accuracy when $p < 0.0001$.) In parallel, if $p = 0.25$ then "…was not statistically signiffcant" would be replaced by "…had $p = 0.25$," and "$p > 0.05$" would be replaced by "$p = 0.25$."
2. Present p-values for more than one possibility when testing a targeted parameter. For example, if you discuss the p-value from a test of a null hypothesis, also discuss alongside this null p-value another p-value for a plausible alternative parameter possibility (ideally the one used to calculate power in the study proposal). As another example: if you do an equivalence test, present the p-values for both the lower and upper bounds of the equivalence interval (which are used for equivalence tests based on two one-sided tests).
3. Show conffdence intervals for targeted study parameters, but also supplement them with p-values for testing relevant hypotheses (e.g., the p-values for both the null and the alternative hypotheses used for the study design or proposal, as in #2). Conffdence intervals only show clearly what is in or out of the interval (i.e., a 95% interval only shows clearly what has $p > 0.05$ or $p \leq 0.05$), but more detail is offen desirable for key hypotheses under contention.
4. Compare groups and studies directly by showing p-values and interval estimates for their differences, not by comparing p-values or interval estimates from the two groups or studies. For example, seeing $p = 0.03$ in males and $p = 0.12$ in females does ***not*** mean that different associations were seen in males and females; instead, one needs a p-value and conffdence interval for the difference in the sex-speciffc associations to examine the between-sex difference. Similarly, if an early study reported a conffdence interval which excluded the null and then a subsequent study reported a conffdence interval which included the null, that does not mean the studies gave confficting results or that the second study failed to replicate the ffrst study; instead, one needs a p-value and conffdence interval for the difference in the study-speciffc associations to examine the between-study difference. In all cases, differences-between-differences must be analyzed directly by statistics for that purpose.
5. Supplement a focal p-value p with its Shannon information transform (s-value or surprisal) $s = -\log_2(p)$. This measures the amount of information supplied by the test against the tested hypothesis (or model): Rounded off, the s-value s shows the number of heads in a row one would need to see when tossing a coin to get the same amount of information against the tosses being "fair" (independent with "heads" probability of 1/2) instead of being loaded for heads. For example, if $p = 0.03$, this represents $-\log_2(0.03) = 5$ bits of information against the hypothesis (like getting 5 heads in a trial of "fairness" with 5 coin tosses); and if $p = 0.25$, this represents only $-\log_2(0.25) = 2$ bits of information against the hypothesis (like getting 2 heads in a trial of "fairness" with only 2 coin tosses).

Betensky, R., The p-Value Requires Context, Not a Threshold

1. Interpret the p-value in light of its context of sample size and meaningful effect size.
2. Incorporate the sample size and meaningful effect size into a decision to reject the null hypothesis.

Anderson, A., Assessing Statistical Results: Magnitude, Precision and Model Uncertainty

1. Evaluate the importance of statistical results based on their practical implications.
2. Evaluate the strength of empirical evidence based on the precision of the estimates and the plausibility of the modeling choices.
3. Seek out subject matter expertise when evaluating the importance and the strength of empirical evidence.

Heck, P., and Krueger, J., Putting the p-Value in Its Place

1. Use the p-value as a heuristic, that is, as the base for a tentative inference regarding the presence or absence of evidence against the tested hypothesis.
2. Supplement the p-value with other, conceptually distinct methods and practices, such as effect size estimates, likelihood ratios, or graphical representations.
3. Strive to embed statistical hypothesis testing within strong *a priori* theory and a context of relevant prior empirical evidence.

Johnson, V., Evidence From Marginally Significant t-Statistics

1. Be transparent in the number of outcome variables that were analyzed.
2. Report the number (and values) of all test statistics that were calculated.
3. Provide access to protocols for studies involving human or animal subjects.

4. Clearly describe data values that were excluded from analysis and the justiffcation for doing so.
5. Provide suff cient details on experimental design so that other researchers can replicate the experiment.
6. Describe only p-values less than 0.005 as being "statistically signiffcant."

Fraser, D., The p-Value Function and Statistical Inference

1. Determine a primary variable for assessing the hypothesis at issue.
2. Calculate its well deffned distribution function, respecting continuity.
3. Substitute the observed data value to obtain the "p-value function."
4. Extract the available well deffned conffdence bounds, conffdence intervals, and median estimate.
5. Know that you don't have an intellectual basis for decisions.

Rougier, J., p-Values, Bayes Factors, and Sufi ciency

1. Recognize that behind every choice of null distribution and test statistic, there lurks a plausible family of alternative hypotheses, which can provide more insight into the null distribution.

Rose, S., and McGuire, T., Limitations of p-Values and R-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment

1. Formulate a clear objective for variable inclusion in regression procedures.
2. Assess all relevant evaluation metrics.
3. Incorporate algorithmic fairness considerations.

7.3. Supplementing or Replacing p

Blume, J., Greevy, R., Welty, V., Smith, J., and DuPont, W., An Introduction to Second Generation p-Values

1. Construct a composite null hypothesis by specifying the range of effects that are not scientiffcally meaningful (do this before looking at the data). Why: Eliminating the conffict between scientiffc signiffcance and statistical signiffcance has numerous statistical and scientiffc beneffts.
2. Replace classical p-values with second-generation p-values (SGPV). Why: SGPVs accommodate composite null hypotheses and encourage the proper communication of ffndings.
3. Interpret the SGPV as a high-level summary of what the data say. Why: Science needs a simple indicator of when the data support only meaningful effects (SGPV = 0), when the data support only trivially null effects (SGPV = 1), or when the data are inconclusive ($0 < \text{SGPV} < 1$).
4. Report an interval estimate of effect size (conffdence interval, support interval, or credible interval) and note its proximity to the composite null hypothesis. Why: This is a more detailed description of study ffndings.
5. Consider reporting false discovery rates with SGPVs of 0 or 1. Why: FDRs gauge the chance that an inference is incorrect under assumptions about the data generating process and prior knowledge.

Goodman, W., Spruill, S., and Komarofi, E., A Proposed Hybrid Efiect Size Plus p-Value Criterion: Empirical Evidence Supporting Its Use

1. Determine how far the true parameter's value would have to be, in your research context, from exactly equaling the conventional, point null hypothesis to consider that the distance is meaningfully large or practically signiffcant.
2. Combine the conventional p-value criterion with a minimum effect size criterion to generate a two-criteria inference-indicator signal, which provides heuristic, but nondeffnitive evidence, for inferring the parameter's true location.
3. Document the intended criteria for your inference procedures, such as a p-value cut-point and a minimum practically signiffcant effect size, prior to undertaking the procedure.
4. Ensure that you use the appropriate inference method for the data that are obtainable and for the inference that is intended.
5. Acknowledge that every study is fraught with limitations from unknowns regarding true data distributions and other conditions that one's method assumes.

Benjamin, D., and Berger, J., Three Recommendations for Improving the Use of p-Values

1. Replace the 0.05 "statistical signiffcance" threshold for claims of novel discoveries with a 0.005 threshold and refer to p-values between 0.05 and 0.005 as "suggestive."
2. Report the data-based odds of the alternative hypothesis to the null hypothesis. If the data-based odds cannot be calculated, then use the p-value to report an upper bound on the data-based odds: $1/(-ep \ln p)$.
3. Report your prior odds and posterior odds (prior odds * data-based odds) of the alternative hypothesis to the null hypothesis. If the data-based odds cannot be calculated, then use your prior odds and the p-value to report an upper bound on your posterior odds: (prior odds) * $(1/(-ep \ln p))$.

Colquhoun, D., The False Positive Risk: A Proposal Concerning What to Do About p-Values

1. Continue to provide p-values and conffdence intervals. Although widely misinterpreted, people know how to calculate them and they aren't entirely useless. Just don't ever use the terms "statistically signiffcant" or "nonsigniffcant."
2. Provide in addition an indication of false positive risk (FPR). This is the probability that the claim of a real effect on the basis of the p-value is in fact false. The FPR (not the p-value) is the probability that your result occurred by chance. For example, the fact that, under plausible assumptions, observation of a p-value close to 0.05 corresponds to an FPR of at least 0.2–0.3 shows clearly the weakness of the conventional criterion for "statistical signiffcance."
3. Alternatively, specify the prior probability of there being a real effect that one would need to be able to justify in order to achieve an FPR of, say, 0.05.

Notes:

There are many ways to calculate the FPR. One, based on a point null and simple alternative can be calculated with the web calculator at *http://fpr-calc.ucl.ac.uk/*. However other approaches to the calculation of FPR, based on different

assumptions, give results that are similar (Table 1 in Colquhoun 2019).

To calculate FPR it is necessary to specify a prior probability and this is rarely known. My recommendation 2 is based on giving the FPR for a prior probability of 0.5. Any higher prior probability of there being a real effect is not justifiable in the absence of hard data. In this sense, the calculated FPR is the minimum that can be expected. More implausible hypotheses would make the problem worse. For example, if the prior probability of there being a real effect were only 0.1, then observation of $p = 0.05$ would imply a disastrously high FPR = 0.76, and in order to achieve an FPR of 0.05, you'd need to observe $p = 0.00045$. Others (especially Goodman) have advocated giving likelihood ratios (LRs) in place of p-values. The FPR for a prior of 0.5 is simply $1/(1 + \text{LR})$, so to give the FPR for a prior of 0.5 is simply a more-easily-comprehensible way of specifying the LR, and so should be acceptable to frequentists and Bayesians.

Matthews, R., Moving Toward the Post p < 0.05 Era via the Analysis of Credibility

1. Report the outcome of studies as effect sizes summarized by confidence intervals (CIs) along with their point estimates.
2. Make full use of the point estimate and width and location of the CI relative to the null effect line when interpreting findings. The point estimate is generally the effect size best supported by the study, irrespective of its statistical significance/nonsignificance. Similarly, tight CIs located far from the null effect line generally represent more compelling evidence for a nonzero effect than wide CIs lying close to that line.
3. Use the analysis of credibility (AnCred) to assess quantitatively the credibility of inferences based on the CI. AnCred determines the level of prior evidence needed for a new finding to provide credible evidence for a nonzero effect.
4. Establish whether this required level of prior evidence is supported by current knowledge and insight. If it is, the new result provides credible evidence for a nonzero effect, irrespective of its statistical significance/nonsignificance.

Gannon, M., Pereira, C., and Polpo, A., Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels

1. Retain the useful concept of statistical significance and the same operational procedures as currently used for hypothesis tests, whether frequentist (Neyman–Pearson p-value tests) or Bayesian (Bayes-factor tests).
2. Use tests with a sample-size-dependent significance level—ours is optimal in the sense of the generalized Neyman–Pearson lemma.
3. Use a testing scheme that allows tests of any kind of hypothesis, without restrictions on the dimensionalities of the parameter space or the hypothesis. Note that this should include "sharp" hypotheses, which correspond to subsets of lower dimensionality than the full parameter space.
4. Use hypothesis tests that are compatible with the likelihood principle (LP). They can be easier to interpret consistently than tests that are not LP-compliant.
5. Use numerical methods to handle hypothesis-testing problems with high-dimensional sample spaces or parameter spaces.

Pogrow, S., How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings

1. Switch from reliance on statistical or practical significance to the more stringent statistical criterion of practical benefit for (a) assessing whether applied research findings indicate that an intervention is effective and should be adopted and scaled—particularly in complex organizations such as schools and hospitals and (b) determining whether relationships are sufficiently strong and explanatory to be used as a basis for setting policy or practice recommendations. Practical benefit increases the likelihood that observed benefits will replicate in subsequent research and in clinical practice by avoiding the problems associated with relying on small effect sizes.
2. Reform statistics courses in applied disciplines to include the principles of practical benefit, and have students review influential applied research articles in the discipline to determine which findings demonstrate practical benefit.
3. Recognize the need to develop different inferential statistical criteria for assessing the importance of applied research findings as compared to assessing basic research findings.
4. Consider consistent, noticeable improvements across contexts using the quick prototyping methods of improvement science as a preferable methodology for identifying effective practices rather than on relying on RCT methods.
5. Require that applied research reveal the actual unadjusted means/medians of results for all groups and subgroups, and that review panels take such data into account—as opposed to only reporting relative differences between adjusted means/medians. This will help preliminarily identify whether there appear to be clear benefits for an intervention.

7.4. Adopting More Holistic Approaches

McShane, B., Gal, D., Gelman, A., Robert, C., and Tackett, J., Abandon Statistical Significance

1. Treat p-values (and other purely statistical measures like confidence intervals and Bayes factors) continuously rather than in a dichotomous or thresholded manner. In doing so, bear in mind that it seldom makes sense to calibrate evidence as a function of p-values or other purely statistical measures because they are, among other things, typically defined relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error.
2. Give consideration to related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain. Do this always—not just once some p-value or other statistical threshold has been attained—and do this without giving priority to p-values or other purely statistical measures.

3. Analyze and report all of the data and relevant results rather than focusing on single comparisons that attain some p-value or other statistical threshold.
4. Conduct a decision analysis: p-value and other statistical threshold-based rules implicitly express a particular tradeoff between Type I and Type II error, but in reality this tradeoff should depend on the costs, benefits, and probabilities of all outcomes.
5. Accept uncertainty and embrace variation in effects: we can learn much (indeed, more) about the world by forsaking the false promise of certainty offered by dichotomous declarations of truth or falsity—binary statements about there being "an effect" or "no effect"—based on some p-value or other statistical threshold being attained.
6. Obtain more precise individual-level measurements, use within-person or longitudinal designs more often, and give increased consideration to models that use informative priors, that feature varying treatment effects, and that are multilevel or meta-analytic in nature.

Tong, C., Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science

1. Prioritize effort for sound data production: the planning, design, and execution of the study.
2. Build scientific arguments with many sets of data and multiple lines of evidence.
3. Recognize the difference between exploratory and confirmatory objectives and use distinct statistical strategies for each.
4. Use flexible descriptive methodology, including disciplined data exploration, enlightened data display, and regularized, robust, and nonparametric models, for exploratory research.
5. Restrict statistical inferences to confirmatory analyses for which the study design and statistical analysis plan are prespecified prior to, and strictly adhered to during, data acquisition.

Amrhein, V., Trafimow, D., and Greenland, S., Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don't Expect Replication

1. Do not dichotomize, but embrace variation.
   (a) Report and interpret inferential statistics like the p-value in a continuous fashion; do not use the word "significant."
   (b) Interpret interval estimates as "compatibility intervals," showing effect sizes most compatible with the data, under the model used to compute the interval; do not focus on whether such intervals include or exclude zero.
   (c) Treat inferential statistics as highly unstable local descriptions of relations between models and the obtained data.
      (i) Free your "negative results" by allowing them to be potentially positive. Most studies with large p-values or interval estimates that include the null should be considered "positive," in the sense that they usually leave open the possibility of important effects (e.g., the effect sizes within the interval estimates).
      (ii) Free your "positive results" by allowing them to be different. Most studies with small p-values or interval estimates that are not near the null should be considered provisional, because in replication studies the p-values could be large and the interval estimates could show very different effect sizes.
      (iii) There is no replication crisis if we don't expect replication. Honestly reported results *must* vary from replication to replication because of varying assumption violations and random variation; excessive agreement itself would suggest deeper problems such as failure to publish results in conflict with group expectations.

Calin-Jageman, R., and Cumming, G., The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known

1. Ask quantitative questions and give quantitative answers.
2. Countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error.
3. Seek replication, and use quantitative methods to synthesize across data sets as a matter of course.
4. Use Open Science practices to enhance the trustworthiness of research results.
5. Avoid, wherever possible, any use of p-values or NHST.

Ziliak, S., How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little "p" Is Not Enough

- *G-10 Consider the Purpose of the Inquiry, and Compare with Best Practice.* Falsification of a null hypothesis is not the main purpose of the experiment or observational study. Making money or beer or medicine—ideally more and better than the competition and best practice—is. Estimating the importance of your coefficient relative to results reported by others, is. To repeat, as the 2016 ASA Statement makes clear, merely falsifying a null hypothesis with a qualitative yes/no, exists/does not exist, significant/not significant answer, is not itself significant science, and should be eschewed.
- *G-9 Estimate the Stakes (Or Eat Them).* Estimation of magnitudes of effects, and demonstrations of their substantive meaning, should be the center of most inquiries. Failure to specify the stakes of a hypothesis is the first step toward eating them (gulp).
- *G-8 Study Correlated Data: ABBA, Take a Chance on Me.* Most regression models assume "iid" error terms—independently and identically distributed—yet most data in the social and life sciences are correlated by systematic, nonrandom effects—and are thus not independent. Gosset solved the problem of correlated soil plots with the "ABBA" layout, maximizing the correlation of paired differences between the As and Bs with a perfectly balanced chiasmic arrangement.
- *G-7 Minimize "Real Error" with the 3 R's: Represent, Replicate, Reproduce.* A test of significance on a single set of data is nearly valueless. Fisher's p, Student's t, and other tests should only be used when there is actual repetition of the experi-

ment. "One and done" is scientism, not scientiffc. Random error is not equal to real error, and is usually smaller and less important than the sum of nonrandom errors. Measurement error, confounding, speciffcation error, and bias of the auspices are frequently larger in all the testing sciences, agronomy to medicine. Guinnessometrics minimizes real error by repeating trials on stratiffed and balanced yet independent experimental units, controlling as much as possible for local ffxed effects.

- *G-6 Economize with "Less is More": Small Samples of Independent Experiments*. Small sample analysis and distribution theory has an economic origin and foundation: changing inputs to the beer on the large scale (for Guinness, enormous global scale) is risky, with more than money at stake. But smaller samples, as Gosset showed in decades of barley and hops experimentation, does not mean "less than," and Big Data is in any case not the solution for many problems.
- *G-5 Keep Your Eyes on the Size Matters/How Much? Question*. There will be distractions but the expected loss and profft functions rule, or should. Are regression coeff cients or differences between means large or small? Compared to what? How do you know?
- *G-4 Visualize*. Parameter uncertainty is not the same thing as model uncertainty. Does the result hit you between the eyes? Does the study show magnitudes of effects across the entire distribution? Advances in visualization soffware continue to outstrip advances in statistical modeling, making more visualization a no brainer.
- *G-3 Consider Posteriors and Priors too ("It pays to go Bayes")*. The sample on hand is rarely the only thing that is "known." Subject matter expertise is an important prior input to statistical design and affects analysis of "posterior" results. For example, Gosset at Guinness was wise to keep quality assurance metrics and bottom line profft at the center of his inquiry. How does prior information fft into the story and evidence? Advances in Bayesian computing soffware make it easier and easier to do a Bayesian analysis, merging prior and posterior information, values, and knowledge.
- *G-2 Cooperate Up, Down, and Across (Networks and Value Chains)*. For example, where would brewers be today without the continued cooperation of farmers? Perhaps back on the farm and not at the brewery making beer. Statistical science is social, and cooperation helps. Guinness ffnanced a large share of modern statistical theory, and not only by supporting Gosset and other brewers with academic sabbaticals (Ziliak and McCloskey 2008).
- *G-1 Answer the Brewer's Original Question ("How should you set the odds?")*. No bright-line rule of statistical signiffcance can answer the brewer's question. As Gosset said way back in 1904, how you set the odds depends on "the importance of the issues at stake" (e.g., the expected benefft and cost) together with the cost of obtaining new material.

Billheimer, D., Predictive Inference and Scientific Reproducibility

1. Predict observable events or quantities that you care about.
2. Quantify the uncertainty of your predictions.

Manski, C., Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing

1. Statisticians should relearn statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten by the century's end.
2. Statistical decision theory should supplant hypothesis testing when statisticians study treatment choice with trial data.
3. Statisticians should use statistical decision theory when analyzing decision making with sample data more generally.

Manski, C., and Tetenov, A., Trial Size for Near Optimal Choice between Surveillance and Aggressive Treatment: Reconsidering MSLT-II

1. Statisticians should relearn statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten by the century's end.
2. Statistical decision theory should supplant hypothesis testing when statisticians study treatment choice with trial data.
3. Statisticians should use statistical decision theory when analyzing decision making with sample data more generally.

Lavine, M., Frequentist, Bayes, or Other?

1. Look for and present results from many models that fft the data well.
2. Evaluate models, not just procedures.

Ruberg, S., Harrell, F., Gamalo-Siebers, M., LaVange, L., Lee J., Price K., and Peck C., Inference and Decision-Making for 21st Century Drug Development and Approval

1. Apply Bayesian paradigm as a framework for improving statistical inference and regulatory decision making by using probability assertions about the magnitude of a treatment effect.
2. Incorporate prior data and available information formally into the analysis of the confffrmatory trials.
3. Justify and pre-specify how priors are derived and perform sensitivity analysis for a better understanding of the impact of the choice of prior distribution.
4. Employ quantitative utility functions to reffect key considerations from all stakeholders for optimal decisions via a probability-based evaluation of the treatment effects.
5. Intensify training in Bayesian approaches, particularly for decision makers and clinical trialists (e.g., physician scientists in FDA, industry and academia).

van Dongen, N., Wagenmakers, E.J., van Doorn, J., Gronau, Q., van Ravenzwaaij, D., Hoekstra, R., Haucke, M., Lakens, D., Hennig, C., Morey, R., Homer, S., Gelman, A., and Sprenger, J., Multiple Perspectives on Inference for Two Simple Statistical Scenarios

1. Clarify your statistical goals explicitly and unambiguously.
2. Consider the question of interest and choose a statistical approach accordingly.
3. Acknowledge the uncertainty in your statistical conclusions.
4. Explore the robustness of your conclusions by executing several different analyses.
5. Provide enough background information such that other researchers can interpret your results and possibly execute meaningful alternative analyses.

7.5. Reforming Institutions: Changing Publication Policies and Statistical Education

Trafimow, D., Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post P < 0.05 Universe

1. Tolerate ambiguity.
2. Replace signiffcance testing with a priori thinking.
3. Consider the nature of the contribution, on multiple levels.
4. Emphasize thinking and execution, not results.
5. Consider that the assumption of random and independent sampling might be wrong.

Locascio, J., The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration

For journal reviewers

1. Provide an initial provisional decision regarding acceptance for publication of a journal manuscript based exclusively on the judged importance of the research issues addressed by the study and the soundness of the reported methodology. (The latter would include appropriateness of data analysis methods.) Give no weight to the reported results of the study per se in the decision as to whether to publish or not.
2. To ensure #1 above is accomplished, commit to an initial decision regarding publication affer having been provided with only the Introduction and Methods sections of a manuscript by the editor, not having seen the Abstract, Results, or Discussion. (The latter would be reviewed only if and affer a generally irrevocable decision to publish has already been made.)

For investigators/manuscript authors

1. Obtain consultation and collaboration from statistical consultant(s) and research methodologist(s) early in the development and conduct of a research study.
2. Emphasize the clinical and scientiffc importance of a study in the Introduction section of a manuscript, and give a clear, explicit statement of the research questions being addressed and any hypotheses to be tested.
3. Include a detailed statistical analysis subsection in the Methods section, which would contain, among other things, a justiffcation of the adequacy of the sample size and the reasons various statistical methods were employed. For example, if null hypothesis signiffcance testing and *p*-values are used, presumably supplemental to other methods, justify why those methods apply and will provide useful additional information in this particular study.
4. Submit for publication reports of well-conducted studies on important research issues regardless of ffndings, for example, even if only null effects were obtained, hypotheses were not conffrmed, mere replication of previous results were found, or results were inconsistent with established theories.

Hurlbert, S., Levine, R., and Utts, J., Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires

1. Encourage journal editorial boards to disallow use of the phrase "statistically signiffcant," or even "signiffcant," in manuscripts they will accept for review.
2. Give primary emphasis in abstracts to the magnitudes of those effects most conclusively demonstrated and of greatest import to the subject matter.
3. Report precise *p*-values or other indices of evidence against null hypotheses as continuous variables not requiring any labeling.
4. Understand the meaning of and rationale for neoFisherian signiffcance assessment (NFSA).

Campbell, H., and Gustafson, P., The World of Research Has Gone Berserk: Modeling the Consequences of Requiring "Greater Statistical Stringency" for Scientific Publication

1. *Consider the meta-research implications of implementing new publication/funding policies.* Journal editors and research funders should attempt to model the impact of proposed policy changes before any implementation. In this way, we can anticipate the policy impacts (both positive and negative) on the types of studies researchers pursue and the types of scientiffc articles that ultimately end up published in the literature.

Fricker, R., Burke, K., Han, X., and Woodall, W., Assessing the Statistical Analyses Used in Basic and Applied Social Psychology Afier Their p-Value Ban

1. Use measures of statistical signiffcance combined with measures of practical signiffcance, such as conffdence intervals on effect sizes, in assessing research results.
2. Classify research results as either exploratory or conffrmatory and appropriately describe them as such in all published documentation.
3. Deffne precisely the population of interest in research studies and carefully assess whether the data being analyzed are representative of the population.
4. Understand the limitations of inferential methods applied to observational, convenience, or other nonprobabilistically sampled data.

Maurer, K., Hudiburgh, L., Werwinski, L., and Bailer J., Content Audit for p-Value Principles in Introductory Statistics

1. Evaluate the coverage of *p*-value principles in the introductory statistics course using rubrics or other systematic assessment guidelines.
2. Discuss and deploy improvements to curriculum coverage of *p*-value principles.
3. Meet with representatives from other departments, who have majors taking your statistics courses, to make sure that inference is being taught in a way that ffts the needs of their disciplines.
4. Ensure that the correct interpretation of *p*-value principles is a point of emphasis for all faculty members and embedded within all courses of instruction.

Steel, A., Liermann, M., and Guttorp, P., Beyond Calculations: A Course in Statistical Thinking

1. Design curricula to teach students how statistical analyses are embedded within a larger science life-cycle, including steps such as project formulation, exploratory graphing, peer review, and communication beyond scientists.
2. Teach the *p*-value as only one aspect of a complete data analysis.

3. Prioritize helping students build a strong understanding of what testing and estimation can tell you over teaching statistical procedures.
4. Explicitly teach statistical communication. Effective communication requires that students clearly formulate the benefits and limitations of statistical results.
5. Force students to struggle with poorly defined questions and real, messy data in statistics classes.

    5. Encourage students to match the mathematical metric (or data summary) to the scientific question. Teaching students to create customized statistical tests for custom metrics allows statistics to move beyond the mean and pinpoint specific scientific questions.

Acknowledgments

Without the help of a huge team, this special issue would never have happened. The articles herein are about the equivalent of three regular issues of *The American Statistician*. Thank you to all the authors who submitted papers for this issue. Thank you, authors whose papers were accepted, for enduring our critiques. We hope they made you happier with your finished product. Thank you to a talented, hard-working group of associate editors for handling many papers: Frank Bretz, George Cobb, Doug Hubbard, Ray Hubbard, Michael Lavine, Fan Li, Xihong Lin, Tom Louis, Regina Nuzzo, Jane Pendergast, Annie Qu, Sherri Rose, and Steve Ziliak. Thank you to all who served as reviewers. We definitely couldn't have done this without you. Thank you, TAS Editor Dan Jeske, for your vision and your willingness to let us create this special issue. Special thanks to Janet Wallace, TAS editorial coordinator, for spectacular work and tons of patience. We also are grateful to ASA Journals Manager Eric Sampson for his leadership, and to our partners, the team at Taylor and Francis, for their commitment to ASA's publishing efforts. Thank you to all who read and commented on the draft of this editorial. You made it so much better! Regina Nuzzo provided extraordinarily helpful substantive and editorial comments. And thanks most especially to the ASA Board of Directors, for generously and enthusiastically supporting the "*p*-values project" since its inception in 2014. Thank you for your leadership of our profession and our association.

Gratefully,
Ronald L. Wasserstein
American Statistical Association, Alexandria, VA
ron@amstat.org

Allen L. Schirm
Mathematica Policy Research (retired), Washington, DC
allenschirm@gmail.com

Nicole A. Lazar
Department of Statistics, University of Georgia, Athens, GA
nlazar@stat.uga.edu

References

References to articles in this special issue

Amrhein, V., Traffmow, D., and Greenland, S. (2019), "Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don't Expect Replication," *The American Statistician*, 73. [2,3,4,5,6,7,8,9]

Anderson, A. (2019), "Assessing Statistical Results: Magnitude, Precision and Model Uncertainty," *The American Statistician*, 73. [3]

Benjamin, D., and Berger, J. (2019), "Three Recommendations for Improving the Use of *p*-Values," *The American Statistician*, 73. [5,7,9]

Betensky, R. (2019), "The *p*-Value Requires Context, Not a Threshold," *The American Statistician*, 73. [4,6]

Billheimer, D. (2019), "Predictive Inference and Scientiffc Reproducibility," *The American Statistician*, 73. [5]

Blume, J., Greevy, R., Welty, V., Smith, J., and DuPont, W. (2019), "An Introduction to Second Generation *p*-Value," *The American Statistician*, 73. [4]

Brownstein, N., Louis, T., O'Hagan, A., and Pendergast, J. (2019), "The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making," *The American Statistician*, 73. [5]

Calin-Jageman, R., and Cumming, G. (2019), "The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known," *The American Statistician*, 73. [3,5,9,10]

Campbell, H., and Gustafson, P. (2019), "The World of Research Has Gone Berserk: Modeling the Consequences of Requiring 'Greater Statistical Stringency' for Scientiffc Publication," *The American Statistician*, 73. [8]

Colquhoun, D. (2019), "The False Positive Risk: A Proposal Concerning What to Do About *p*-Value," *The American Statistician*, 73. [4,7,14]

Fraser, D. (2019), "The *p*-Value Function and Statistical Inference," *The American Statistician*, 73. [6]

Fricker, R., Burke, K., Han, X., and Woodall, W (2019), "Assessing the Statistical Analyses Used in Basic and Applied Social Psychology Affer Their *p*-Value Ban," *The American Statistician*, 73. [7,9]

Gannon, M., Pereira, C., and Polpo, A. (2019), "Blending Bayesian and Classical Tools to Deffne Optimal Sample-Size-Dependent Signiffcance Levels," *The American Statistician*, 73. [5]

Goodman, S. (2019), "Why is Getting Rid of *p*-Values So Hard? Musings on Science and Statistics," *The American Statistician*, 73. [7,8,10]

Goodman, W., Spruill, S., and Komaroff, E. (2019), "A Proposed Hybrid Effect Size Plus *p*-Value Criterion: Empirical Evidence Supporting Its Use," *The American Statistician*, 73. [5]

Greenland, S. (2019), "Valid *p*-Values Behave Exactly as They Should: Some Misleading Criticisms of *p*-Values and Their Resolution With *s*-Values," *The American Statistician*, 73. [3,4,5,6,7]

Heck, P., and Krueger, J. (2019), "Putting the *p*-Value in Its Place," *The American Statistician*, 73. [4]

Hubbard, D., and Carriquiry, A. (2019), "Quality Control for Scientiffc Research: Addressing Reproducibility, Responsiveness and Relevance," *The American Statistician*, 73. [5]

Hubbard, R. (2019), "Will the ASA's Efforts to Improve Statistical Practice Be Successful? Some Evidence to the Contrary," *The American Statistician*, 73. [8]

Hubbard, R., Haig, B. D., and Parsa, R. A. (2019), "The Limited Role of Formal Statistical Inference in Scientiffc Inference," *The American Statistician*, 73. [2,7]

Hurlbert, S., Levine, R., and Utts, J. (2019), "Coup de Grâce for a Tough Old Bull: 'Statistically Signiffcant' Expires," *The American Statistician*, 73. [8,9]

Ioannidis, J. (2019), "What Have We (Not) Learnt From Millions of Scientiffc Papers With *p*-Values?," *The American Statistician*, 73. [6]

Johnson, V. (2019), "Evidence From Marginally Signiffcant *t* Statistics," *The American Statistician*, 73. [7]

Kennedy-Shaffer, L. (2019), "Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize *p*-Values and Signiffcance Testing," *The American Statistician*, 73. [7]

Kmetz, J. (2019), "Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of *p*-Values," *The American Statistician*, 73. [8]

Lavine, M. (2019), "Frequentist, Bayes, or Other?," *The American Statistician*, 73. [6]

Locascio, J. (2019), "The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration," *The American Statistician*, 73. [4,5,8,9]

Manski, C. (2019), "Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing," *The American Statistician*, 73. [5]

Manski, C., and Tetenov, A. (2019), "Trial Size for Near Optimal Choice between Surveillance and Aggressive Treatment: Reconsidering MSLT-II," *The American Statistician*, 73. [5]

Matthews, R. (2019), "Moving Toward the Post $p < 0.05$ Era Via the Analysis of Credibility," *The American Statistician*, 73. [4,9]

Maurer, K., Hudiburgh, L., Werwinski, L., and Bailer, J. (2019), "Content Audit for *P*-Value Principles in Introductory Statistics," *The American Statistician*, 73. [8]

McShane, B., Gal, D., Gelman, A., Robert, C., and Tackett, J. (2019), "Abandon Statistical Signiffcance," *The American Statistician*, 73. [4,6,7]

McShane, B., Tackett, J., Böckenholt, U., and Gelman, A. (2019), "Large Scale Replication Projects in Contemporary Psychological Research," *The American Statistician*, 73. [9]

O'Hagan, A. (2019), "Expert Knowledge Elicitation: Subjective But Scientiffc," *The American Statistician*, 73. [5,6]

Pogrow, S. (2019), "How Effect Size (Practical Signiffcance) Misleads Clinical Practice: The Case for Switching to Practical Benefft to Assess Applied Research Findings," *The American Statistician*, 73. [4]

Rose, S., and McGuire, T. (2019), "Limitations of *p*-Values and *R*-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment," *The American Statistician*, 73. [7]

Rougier, J. (2019), "*p*-Values, Bayes Factors, and Suff ciency," *The American Statistician*, 73. [6]

Ruberg, S., Harrell, F., Gamalo-Siebers, M., LaVange, L., Lee J., Price K., and Peck C. (2019), "Inference and Decision-Making for 21st Century Drug Development and Approval," *The American Statistician*, 73. [10]

Steel, A., Liermann, M., and Guttorp, P. (2019), "Beyond Calculations: A Course in Statistical Thinking," *The American Statistician*, 73. [8]

Traffmow, D. (2019), "Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post $p < .05$ Universe," *The American Statistician*, 73. [7]

Tong, C. (2019), "Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science," *The American Statistician*, 73. [2,3,4,6,8,9]

van Dongen, N., Wagenmakers, E. J., van Doorn, J., Gronau, Q., van Ravenzwaaij, D., Hoekstra, R., Haucke, M., Lakens, D., Hennig, C., Morey, R., Homer, S., Gelman, A., and Sprenger, J. (2019), "Multiple Perspectives on Inference for Two Simple Statistical Scenarios," *The American Statistician*, 73. [6]

Ziliak, S. (2019), "How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little 'P' is Not Enough," *The American Statistician*, 73. [2,3]

Other articles or books referenced

Boring, E. G. (1919), "Mathematical vs. Scientiffc Signiffcance," *Psychological Bulletin*, 16, 335–338. [2]

Cumming, G. (2014), "The New Statistics: Why and How," *Psychological Science*, 25, 7–29. [3]

Davidian, M., and Louis, T. (2012), "Why Statistics?" *Science*, 336, 12. [2]

Edgeworth, F. Y. (1885), "Methods of Statistics," *Journal of the Statistical Society of London*, Jubilee Volume, 181–217. [2]

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd. [2]

Gelman, A. (2015), "Statistics and Research Integrity," *European Science Editing*, 41, 13–14. [8]

——— (2016), "The Problems With *p*-Values Are Not Just With *p*-Values," *The American Statistician*, supplemental materials to *ASA Statement on p-Values and Statistical Significance*, 70, 1–2. [3]

Gelman, A., and Hennig, C. (2017), "Beyond Subjective and Objective in Statistics," *Journal of the Royal Statistical Society*, Series A, 180, 967–1033. [5]

Gelman, A., and Stern, H. (2006), "The Difference Between 'Signiffcant' and 'Not Signiffcant' Is Not Itself Statistically Signiffcant," *The American Statistician*, 60, 328–331. [2]

Ghose, T. (2013), "'Just a Theory': 7 Misused Science Words," *Scientific American* (online), available at *https://www.scientificamerican.com/article/just-a-theory-7-misused-science-words/*. [2]

Goodman, S. (2018), "How Sure Are You of Your Result? Put a Number on It," *Nature*, 564. [5]

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Thousand Oaks, CA: Sage. [1]

Junger, S. (1997), *The Perfect Storm: A True Story of Men Against the Sea*, New York: W.W. Norton. [10]

Locascio, J. (2017), "Results Blind Science Publishing," *Basic and Applied Social Psychology*, 39, 239–246. [8]

Mayo, D. (2018), "Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars," Cambridge, UK: University Printing House. [1]

McShane, B., and Gal, D. (2016), "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence," *Management Science*, 62, 1707–1718. [9]

——— (2017), "Statistical Signiffcance and the Dichotomization of Evidence." *Journal of the American Statistical Association*, 112, 885–895. [9]

Mogil, J. S., and Macleod, M. R. (2017), "No Publication Without Confirmation," *Nature*, 542, 409–411, available at *https://www.nature.com/news/no-publication-without-confirmation-1.21509*. [8]

Rosenthal, R. (1979), "File Drawer Problem and Tolerance for Null Results," *Psychological Bulletin* 86, 638–641. [2]

Wasserstein, R., and Lazar, N. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [1]

Wellek, S. (2017), "A Critical Evaluation of the Current *p*-Value Controversy" (with discussion), *Biometrical Journal*, 59, 854–900. [4,9]

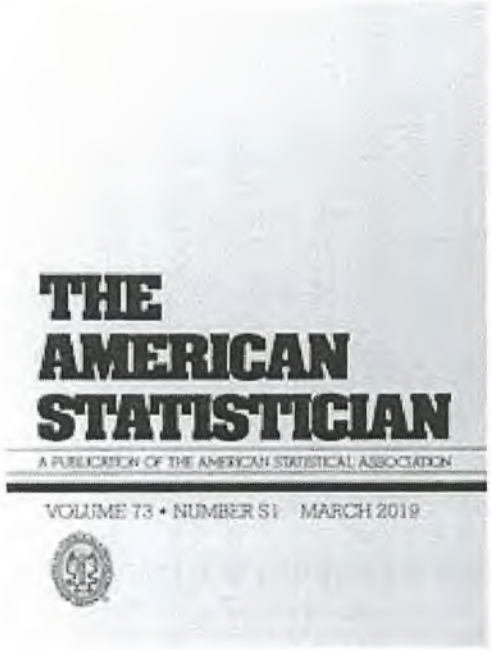
Ziliak, S., and McCloskey, D. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [1,16]

Exhibit 62

JOHNS HOPKINS
INSTITUTE for CLINICAL & TRANSLATIONAL RESEARCH

THE AMERICAN STATISTICIAN SPECIAL ISSUE: MOVING TO A WORLD BEYOND "P < 0.05"

Posted by: Crystal Williams on: March 25, 2019 | Print This Page



"Moving to a World Beyond 'p<0.05'", in a special issue of The American Statistician (*TAS*), the lead editorial calls for abandoning the use of "statistically significant," and offers much (not just one thing) to replace it. Written by Ron Wasserstein, Allen Schirm, and Nicole Lazar, the co-editors of the special issue, summarizes the content of the issue's 43 articles.

These articles discuss the use of p-values and statistical significance that Johns Hopkins researchers may find beneficial.

To read the full article, click here.

750 E. Pratt Street, 16th Floor
Baltimore, MD 21202
410.361.7880 | ictr@jhmi.edu |

The ICTR is funded by the National Center for Advancing Translational Sciences (NCATS) through the Clinical & Translational Science Awards Program

JOHNS HOPKINS
INSTITUTE for CLINICAL & TRANSLATIONAL RESEARCH

ABOUT THE ICTR

The Johns Hopkins Institute for Clinical and Translational Research (ICTR), established in 2007, is one of more than 60 medical research institutions working together as a national consortium to improve the way biomedical research is conducted across the country.

The ICTR addresses obstacles in translating basic science discoveries into research in humans, translating clinical discoveries into the community and communicating experience from clinical practice back to researchers. The ICTR houses three Translational Research Communities for investigators across multiple disciplines that focus on drugs, biologics, vaccines and devices; biomarkers and diagnostic tests; and behavioral, social and systems interventions. These communities of researchers help prioritize clinical problems in need of new treatments, apply new technologies and methodologies, support junior investigators, work with translational partners outside of Johns Hopkins, fund pilot projects, provide regulatory assistance and promote efficient research. Another ICTR program, The Research Studio, provides both a place and a process for investigators and their teams to obtain multidisciplinary guidance to solve clinical and translational research problems.

Through a robust portfolio of training, education and career development programs, the ICTR also provides rigorous, comprehensive training to medical students, graduate students, fellows, junior faculty, practicing physicians and the wider research team, thus promoting the most effective, efficient, collaborative translational research enterprise.

Finally, using the Accelerating Translational Incubator Pilot grants, the ICTR encourages new translational research teams to take the risks necessary to go beyond their usual expertise and to find new collaborators, seek out new partners inside and outside the academic center, and learn new skills necessary to create the interventions the

public is expecting.

ANNUAL REPORTS

ICTR LEADERSHIP

ICTR KICKOFF MEETING (VIDEO)

RESOURCES AND ENVIRONMENT

GRANT INFORMATION

ICTR TEMPLATES AND LOGOS

ORGANIZATIONAL CHART

CITING THE GRANT

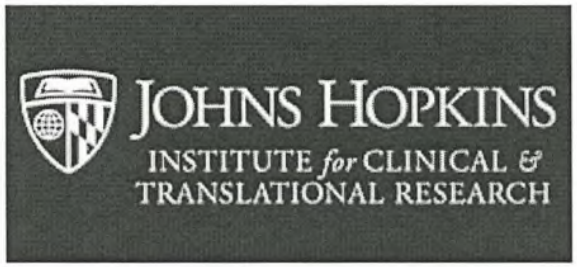
CLINICAL AND TRANSLATIONAL SCIENCE AWARDS (CTSA)

NATIONAL CENTER FOR ADVANCING TRANSLATIONAL SCIENCES (NCATS)

NATIONAL INSTITUTES OF HEALTH

750 E. Pratt Street, 16th Floor
Baltimore, MD 21202
410.361.7880 | ictr@jhmi.edu |

The ICTR is funded by the National Center for Advancing Translational Sciences (NCATS) through the Clinical & Translational Science Awards Program



ICTR LEADERSHIP

Director

DANIEL E. FORD, MD, MPH
Vice Dean for Clinical Investigation
Johns Hopkins School of Medicine

Deputy Directors

CHARLES FLEXNER, MD
Chief Scientific Officer for Strategy and Integration

KAREN BANDEEN-ROCHE, PHD, MS
Deputy Director for Biostatistics and Research Design
Program Director, Biostatistics Center

CHRISTOPHER G. CHUTE, MD, MPH, DrPH
Deputy Director, Informatics Core

CHERYL DENNISON HIMMELFARB, PHD, RN, ANP

Deputy Director for Interdisciplinary Research
Program Director, Research Participant and Community Partnerships

TOBY A. GORDON, SCD
Deputy Director of Business Strategies

CRAIG HENDRIX
Deputy Director, Translational Sciences Core

EDGAR "PETE" MILLER III, MD, PHD
Deputy Director for Clinical Research Education, Training and Career Development

PAMELA OUYANG, MBBS
Deputy Director for the Bayview Medical Center
Associate Program Director, Clinical Research Units

JEFFREY D. ROTHSTEIN MD, PHD
Deputy Director for Pilot Research Initiatives
Director, Accelerated Translational Incubator Pilot (ATIP) Program

ROBERT ALAN WOOD, MD
Deputy Director for Pediatric Research
Associate Program Director, Clinical Research Units

Administrative Staff

ADMINISTRATOR
Mark Garcia

SR. ADMINISTRATIVE MANAGER
Roxanne Stambaugh

COMMUNICATIONS SPECIALIST
Crystal Williams

SR. GRANTS AND CONTRACTS ANALYST
Janet Palmer

SR. ADMINISTRATIVE COORDINATOR
Dawn Childs

BUDGET ANALYST
Jennifer Guntner

750 E. Pratt Street, 16th Floor
Baltimore, MD 21202
410.361.7880 | ictr@jhmi.edu |

The ICTR is funded by the National Center for Advancing Translational Sciences (NCATS) through the Clinical & Translational Science Awards Program

Taylor & Francis

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Moving to a World Beyond "$p < 0.05$"

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p<0.05$", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

To link to this article: https://doi.org/10.1080/00031305.2019.1583913

Published online: 20 Mar 2019.

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913

Taylor & Francis
Taylor & Francis Group

EDITORIAL

Moving to a World Beyond "$p < 0.05$"

Some of you exploring this special issue of *The American Statistician* might be wondering if it's a scolding from pedantic statisticians lecturing you about what *not* to do with p-values, without offering any real ideas of what *to do* about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.

1. "Don't" Is Not Enough

There's not much we can say here about the perils of p-values and significance testing that hasn't been said already for decades (Ziliak and McCloskey 2008; Hubbard 2016). If you're just arriving to the debate, here's a sampling of what not to do:

- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the p-value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

Don't. Don't. Just…don't. Yes, we talk a lot about don'ts. The *ASA Statement on p-Values and Statistical Significance* (Wasserstein and Lazar 2016) was developed primarily because after decades, warnings about the don'ts had gone mostly unheeded. The statement was about what not to do, because there is widespread agreement about the don'ts.

Knowing what not to do with p-values is indeed necessary, but it does not suffice. It is as though statisticians were asking users of statistics to tear out the beams and struts holding up the edifice of modern scientific research without offering solid construction materials to replace them. Pointing out old, rotting timbers was a good start, but now we need more.

Recognizing this, in October 2017, the American Statistical Association (ASA) held the Symposium on Statistical Inference, a two-day gathering that laid the foundations for this special issue of *The American Statistician*. Authors were explicitly instructed to develop papers for the variety of audiences interested in these topics. If you use statistics in research, business, or policymaking but are not a statistician, these articles were indeed written with YOU in mind. And if you are a statistician, there is still much here for you as well.

The papers in this issue propose many new ideas, ideas that in our determination as editors merited publication to enable broader consideration and debate. The ideas in this editorial are likewise open to debate. They are our own attempt to distill the wisdom of the many voices in this issue into an essence of good statistical practice as we currently see it: some do's for teaching, doing research, and informing decisions.

Yet the voices in the 43 papers in this issue do not sing as one. At times in this editorial and the papers you'll hear deep dissonance, the echoes of "statistics wars" still simmering today (Mayo 2018). At other times you'll hear melodies wrapping in a rich counterpoint that may herald an increasingly harmonious new era of statistics. To us, these are all the sounds of statistical inference in the 21st century, the sounds of a world learning to venture beyond "$p < 0.05$."

This is a world where researchers are free to treat "$p = 0.051$" and "$p = 0.049$" as not being categorically different, where authors no longer find themselves constrained to selectively publish their results based on a single magic number. In this world, where studies with "$p < 0.05$" and studies with "$p > 0.05$" are not automatically in conflict, researchers will see their results more easily replicated—and, even when not, they will better understand *why*. As we venture down this path, we will begin to see fewer false alarms, fewer overlooked discoveries, and the development of more customized statistical strategies. Researchers will be free to communicate all their findings in all their glorious uncertainty, knowing their work is to be judged by the quality and effective communication of their science, and not by their p-values. As "statistical significance" is used less, statistical thinking will be used more.

The *ASA Statement on P-Values and Statistical Significance* started moving us toward this world. As of the date of publication of this special issue, the statement has been viewed over 294,000 times and cited over 1700 times—an average of about 11 citations per week since its release. Now we must go further. That's what this special issue of *The American Statistician* sets out to do.

To get to the do's, though, we must begin with one more don't.

2. Don't Say "Statistically Significant"

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

Regardless of whether it was ever useful, a declaration of "statistical significance" has today become meaningless. Made broadly known by Fisher's use of the phrase (1925), Edgeworth's (1885) original intention for statistical significance was simply as a tool to indicate when a result warrants further scrutiny. But that idea has been irretrievably lost. Statistical significance was never meant to imply scientific importance, and the confusion of the two was decried soon after its widespread use (Boring 1919). Yet a full century later the confusion persists.

And so the tool has become the tyrant. The problem is not simply use of the word "significant," although the statistical and ordinary language meanings of the word are indeed now hopelessly confused (Ghose 2013); the term should be avoided for that reason alone. The problem is a larger one, however: using bright-line rules for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making (ASA statement, Principle 3). A label of statistical significance adds nothing to what is already conveyed by the value of p; in fact, this dichotomization of p-values makes matters worse.

For example, no p-value can reveal the plausibility, presence, truth, or importance of an association or effect. Therefore, a label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important. Nor does a label of statistical nonsignificance lead to the association or effect being improbable, absent, false, or unimportant. Yet the dichotomization into "significant" and "not significant" is taken as an imprimatur of authority on these characteristics. In a world without bright lines, on the other hand, it becomes untenable to assert dramatic differences in interpretation from inconsequential differences in estimates. As Gelman and Stern (2006) famously observed, the difference between "significant" and "not significant" is not itself statistically significant.

Furthermore, this false split into "worthy" and "unworthy" results leads to the selective reporting and publishing of results based on their statistical significance—the so-called "file drawer problem" (Rosenthal 1979). And the dichotomized reporting problem extends beyond just publication, notes Amrhein, Trafimow, and Greenland (2019): when authors use p-value thresholds to select which findings to discuss in their papers, "their conclusions and what is reported in subsequent news and reviews will be biased…Such selective attention based on study outcomes will therefore not only distort the literature but will slant published descriptions of study results—biasing the summary descriptions reported to practicing professionals and the general public." For the integrity of scientific publishing and research dissemination, therefore, whether a p-value passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight.

To be clear, the problem is not that of having only two labels. Results should not be trichotomized, or indeed categorized into any number of groups, based on arbitrary p-value thresholds. Similarly, we need to stop using confidence intervals as another means of dichotomizing (based, on whether a null value falls within the interval). And, to preclude a reappearance of this problem elsewhere, we must not begin arbitrarily categorizing other statistical measures (such as Bayes factors).

Despite the limitations of p-values (as noted in Principles 5 and 6 of the ASA statement), however, we are not recommending that the calculation and use of continuous p-values be discontinued. Where p-values are used, they should be reported as continuous quantities (e.g., $p = 0.08$). They should also be described in language stating what the value means in the scientific context. We believe that a reasonable prerequisite for reporting any p-value is the ability to interpret it appropriately. We say more about this in Section 3.3.

To move forward to a world beyond "$p < 0.05$," we must recognize afresh that statistical inference is not—and never has been—equivalent to scientific inference (Hubbard, Haig, and Parsa 2019; Ziliak 2019). However, looking to statistical significance for a marker of scientific observations' credibility has created a guise of equivalency. Moving beyond "statistical significance" opens researchers to the real significance of statistics, which is "the science of learning from data, and of measuring, controlling, and communicating uncertainty" (Davidian and Louis 2012).

In sum, "statistically significant"—don't say it and don't use it.

3. There Are Many Do's

With the don'ts out of the way, we can finally discuss ideas for specific, positive, constructive actions. We have a massive list of them in the seventh section of this editorial! In that section, the authors of all the articles in this special issue each provide their own short set of do's. Those lists, and the rest of this editorial, will help you navigate the substantial collection of articles that follows.

Because of the size of this collection, we take the liberty here of distilling our readings of the articles into a summary of what can be done to move beyond "$p < 0.05$." You will find the rich details in the articles themselves.

What you will NOT find in this issue is one solution that majestically replaces the outsized role that statistical significance has come to play. The statistical community has not yet converged on a simple paradigm for the use of statistical inference in scientific research—and in fact it may never do so. A one-size-fits-all approach to statistical inference is an inappropriate expectation, even after the dust settles from our current remodeling of statistical practice (Tong 2019). Yet solid principles for the use of statistics do exist, and they are well explained in this special issue.

We summarize our recommendations in two sentences totaling seven words: "Accept uncertainty. Be thoughtful, open, and modest." Remember "ATOM."

3.1. Accept Uncertainty

Uncertainty exists everywhere in research. And, just like with the frigid weather in a Wisconsin winter, there are those who will flee from it, trying to hide in warmer havens elsewhere. Others, however, accept and even delight in the omnipresent cold; these are the ones who buy the right gear and bravely take full advantage of all the wonders of a challenging climate. Significance tests and dichotomized p-values have turned many researchers into scientific snowbirds, trying to avoid dealing with uncertainty by escaping to a "happy place" where results are either statistically significant or not. In the real world, data provide a noisy signal. Variation, one of the causes of uncertainty, is everywhere. Exact replication is difficult to achieve. So it is time to get the right (statistical) gear and "move toward a greater acceptance of uncertainty and embracing of variation" (Gelman 2016).

Statistical methods do not rid data of their uncertainty. "Statistics," Gelman (2016) says, "is often sold as a sort of alchemy that transmutes randomness into certainty, an 'uncertainty laundering' that begins with data and concludes with success as measured by statistical significance." To accept uncertainty requires that we "treat statistical results as being much more incomplete and uncertain than is currently the norm" (Amrhein, Trafimow, and Greenland 2019). We must "countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error" (Calin-Jageman and Cumming 2019).

"Accept uncertainty and embrace variation in effects," advise McShane et al. in Section 7 of this editorial. "[W]e can learn much (indeed, more) about the world by forsaking the false promise of certainty offered by dichotomous declarations of truth or falsity—binary statements about there being 'an effect' or 'no effect'—based on some p-value or other statistical threshold being attained."

We can make acceptance of uncertainty more natural to our thinking by accompanying every point estimate in our research with a measure of its uncertainty such as a standard error or interval estimate. Reporting and interpreting point and interval estimates should be routine. However, simplistic use of confidence intervals as a measurement of uncertainty leads to the same bad outcomes as use of statistical significance (especially, a focus on whether such intervals include or exclude the "null hypothesis value"). Instead, Greenland (2019) and Amrhein, Trafimow, and Greenland (2019) encourage thinking of confidence intervals as "compatibility intervals," which use p-values to show the effect sizes that are most compatible with the data under the given model.

How will **accepting uncertainty** change anything? To begin, it will prompt us to seek better measures, more sensitive designs, and larger samples, all of which increase the rigor of research. It also helps us **be modest** (the fourth of our four principles, on which we will expand in Section 3.4) and encourages "meta-analytic thinking" (Cumming 2014). Accepting uncertainty as inevitable is a natural antidote to the seductive certainty falsely promised by statistical significance. With this new outlook, we will naturally seek out replications and the integration of evidence through meta-analyses, which usually requires point and interval estimates from contributing studies. This will in turn give us more precise overall estimates for our effects and associations. And this is what will lead to the best research-based guidance for practical decisions.

Accepting uncertainty leads us to **be thoughtful**, the second of our four principles.

3.2. Be Thoughtful

What do we mean by this exhortation to "be thoughtful"? Researchers already clearly put much thought into their work. We are not accusing anyone of laziness. Rather, we are envisioning a sort of "statistical thoughtfulness." In this perspective, statistically **thoughtful researchers** begin above all else with clearly expressed objectives. They recognize when they are doing exploratory studies and when they are doing more rigidly pre-planned studies. They invest in producing solid data. They consider not one but a multitude of data analysis techniques. And they think about so much more.

3.2.1. Thoughtfulness in the Big Picture

"[M]ost scientific research is exploratory in nature," Tong (2019) contends. "[T]he design, conduct, and analysis of a study are necessarily flexible, and must be open to the discovery of unexpected patterns that prompt new questions and hypotheses. In this context, statistical modeling can be exceedingly useful for elucidating patterns in the data, and researcher degrees of freedom can be helpful and even essential, though they still carry the risk of overfitting. The price of allowing this flexibility is that the validity of any resulting statistical inferences is undermined."

Calin-Jageman and Cumming (2019) caution that "in practice the dividing line between planned and exploratory research can be difficult to maintain. Indeed, exploratory findings have a slippery way of 'transforming' into planned findings as the research process progresses." At the bottom of that slippery slope one often finds results that don't reproduce.

Anderson (2019) proposes three questions **thoughtful researchers** asked thoughtful researchers evaluating research results: What are the practical implications of the estimate? How precise is the estimate? And is the model correctly specified? The latter question leads naturally to three more: Are the modeling assumptions understood? Are these assumptions valid? And do the key results hold up when other modeling choices are made? Anderson further notes, "Modeling assumptions (including all the choices from model specification to sample selection and the handling of data issues) should be sufficiently documented so independent parties can critique, and replicate, the work."

Drawing on archival research done at the Guinness Archives in Dublin, Ziliak (2019) emerges with ten "G-values" he believes we all wish to maximize in research. That is, we want large G-values, not small p-values. The ten principles of Ziliak's "Guinnessometrics" are derived primarily from his examination of experiments conducted by statistician William Sealy Gosset while working as Head Brewer for Guinness. Gosset took an economic approach to the logic of uncertainty, preferring balanced designs over random ones and estimation of gambles over bright-line "testing." Take, for example, Ziliak's G-value 10: "Consider purpose of the inquiry, and compare with best

practice," in the spirit of what farmers and brewers must do. The purpose is generally NOT to falsify a null hypothesis, says Ziliak. Ask what is at stake, he advises, and determine what magnitudes of change are humanly or scientifically meaningful in context.

Pogrow (2019) offers an approach based on practical benefit rather than statistical or practical significance. This approach is especially useful, he says, for assessing whether interventions in complex organizations (such as hospitals and schools) are effective, and also for increasing the likelihood that the observed benefits will replicate in subsequent research and in clinical practice. In this approach, "practical benefit" recognizes that reliance on small effect sizes can be as problematic as relying on p-values.

Thoughtful research prioritizes sound data production by putting energy into the careful planning, design, and execution of the study (Tong 2019).

Locascio (2019) urges researchers to be prepared for a new publishing model that evaluates their research based on the importance of the questions being asked and the methods used to answer them, rather than the outcomes obtained.

3.2.2. Thoughtfulness Through Context and Prior Knowledge

Thoughtful research considers the scientific context and prior evidence. In this regard, a declaration of statistical significance is the antithesis of thoughtfulness: it says nothing about practical importance, and it ignores what previous studies have contributed to our knowledge.

Thoughtful research looks ahead to prospective outcomes in the context of theory and previous research. Researchers would do well to ask, *What do we already know, and how certain are we in what we know?* And building on that and on the field's theory, *what magnitudes of differences, odds ratios, or other effect sizes are practically important?* These questions would naturally lead a researcher, for example, to use existing evidence from a literature review to identify specifically the findings that would be practically important for the key outcomes under study.

Thoughtful research includes careful consideration of the definition of a meaningful effect size. As a researcher you should communicate this up front, before data are collected and analyzed. Afterwards is just too late; it is dangerously easy to justify observed results after the fact and to overinterpret trivial effect sizes as being meaningful. Many authors in this special issue argue that consideration of the effect size and its "scientific meaningfulness" is essential for reliable inference (e.g., Blume et al. 2019; Betensky 2019). This concern is also addressed in the literature on equivalence testing (Wellek 2017).

Thoughtful research considers "related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain…without giving priority to p-values or other purely statistical measures" (McShane et al. 2019).

Thoughtful researchers "use a toolbox of statistical techniques, employ good judgment, and keep an eye on developments in statistical and data science," conclude Heck and Krueger (2019), who demonstrate how the p-value can be useful to researchers as a heuristic.

3.2.3. Thoughtful Alternatives and Complements to P-Values

Thoughtful research considers multiple approaches for solving problems. This special issue includes some ideas for supplementing or replacing p-values. Here is a short summary of some of them, with a few technical details:

Amrhein, Trafimow, and Greenland (2019) and Greenland (2019) advise that null p-values should be supplemented with a p-value from a test of a pre-specified alternative (such as a minimal important effect size). To reduce confusion with posterior probabilities and better portray evidential value, they further advise that p-values be transformed into s-values (Shannon information, surprisal, or binary logworth) $s = -\log_2(p)$. This measure of evidence affirms other arguments that the evidence against a hypothesis contained in the p-value is not nearly as strong as is believed by many researchers. The change of scale also moves users away from probability misinterpretations of the p-value.

Blume et al. (2019) offer a "second generation p-value (SGPV)," the characteristics of which mimic or improve upon those of p-values but take practical significance into account. The null hypothesis from which an SGPV is computed is a composite hypothesis representing a range of differences that would be practically or scientifically inconsequential, as in equivalence testing (Wellek 2017). This range is determined in advance by the experimenters. When the SGPV is 1, the data only support null hypotheses; when the SGPV is 0, the data are incompatible with any of the null hypotheses. SGPVs between 0 and 1 are inconclusive at varying levels (maximally inconclusive at or near SGPV = 0.5.) Blume et al. illustrate how the SGPV provides a straightforward and useful descriptive summary of the data. They argue that it eliminates the problem of how classical statistical significance does not imply scientific relevance, it lowers false discovery rates, and its conclusions are more likely to reproduce in subsequent studies.

The "analysis of credibility"(AnCred) is promoted by Matthews (2019). This approach takes account of both the width of the confidence interval and the location of its bounds when assessing weight of evidence. AnCred assesses the credibility of inferences based on the confidence interval by determining the level of prior evidence needed for a new finding to provide credible evidence for a nonzero effect. If this required level of prior evidence is supported by current knowledge and insight, Matthews calls the new result "credible evidence for a non-zero effect," irrespective of its statistical significance/nonsignificance.

Colquhoun (2019) proposes continuing the use of continuous p-values, but only in conjunction with the "false positive risk (FPR)." The FPR answers the question, "If you observe a 'significant' p-value after doing a single unbiased experiment, what is the probability that your result is a false positive?" It tells you what most people mistakenly still think the p-value does, Colquhoun says. The problem, however, is that to calculate the FPR you need to specify the prior probability that an effect is real, and it's rare to know this. Colquhoun suggests that the FPR could be calculated with a prior probability of 0.5, the largest value reasonable to assume in the absence of hard prior data. The FPR found this way is in a sense the minimum false positive risk (mFPR); less plausible hypotheses (prior probabilities below 0.5) would give even bigger FPRs. Colquhoun says, but the

mFPR would be a big improvement on reporting a p-value alone. He points out that p-values near 0.05 are, under a variety of assumptions, associated with minimum false positive risks of 20–30%, which should stop a researcher from making too big a claim about the "statistical significance" of such a result.

Benjamin and Berger (2019) propose a different supplement to the null p-value. The Bayes factor bound (BFB)—which under typically plausible assumptions is the value $1/(-ep \ln p)$—represents the upper bound of the ratio of data-based odds of the alternative hypothesis to the null hypothesis. Benjamin and Berger advise that the BFB should be reported along with the continuous p-value. This is an incomplete step toward revising practice, they argue, but one that at least confronts the researcher with the maximum possible odds that the alternative hypothesis is true—which is what researchers often think they are getting with a p-value. The BFB, like the FPR, often clarifies that the evidence against the null hypothesis contained in the p-value is not nearly as strong as is believed by many researchers.

Goodman, Spruill, and Komaroff (2019) propose a two-stage approach to inference, requiring both a small p-value below a pre-specified level and a pre-specified sufficiently large effect size before declaring a result "significant." They argue that this method has improved performance relative to use of dichotomized p-values alone.

Gannon, Pereira, and Polpo (2019) have developed a testing procedure combining frequentist and Bayesian tools to provide a significance level that is a function of sample size.

Manski (2019) and Manski and Tetenov (2019) urge a return to the use of statistical decision theory, which they say has largely been forgotten. Statistical decision theory is not based on p-value thresholds and readily distinguishes between statistical and clinical significance.

Billheimer (2019) suggests abandoning inference about parameters, which are frequently hypothetical quantities used to idealize a problem. Instead, he proposes focusing on the prediction of future observables, and their associated uncertainty, as a means to improving science and decision-making.

3.2.4. Thoughtful Communication of Confidence

Be thoughtful and clear about the level of confidence or credibility that is present in statistical results.

Amrhein, Trafimow, and Greenland (2019) and Greenland (2019) argue that the use of words like "significance" in conjunction with p-values and "confidence" with interval estimates misleads users into overconfident claims. They propose that researchers think of p-values as measuring the compatibility between hypotheses and data, and interpret interval estimates as "compatibility intervals."

In what may be a controversial proposal, Goodman (2018) suggests requiring "that any researcher making a claim in a study accompany it with their estimate of the chance that the claim is true." Goodman calls this the confidence index. For example, along with stating "This drug is associated with elevated risk of a heart attack, relative risk (RR) = 2.4, $p = 0.03$," Goodman says investigators might add a statement such as "There is an 80% chance that this drug raises the risk, and a 60% chance that the risk is at least doubled." Goodman acknowledges, "Although simple on paper, requiring a confidence index would entail a profound overhaul of scientific and statistical practice."

In a similar vein, Hubbard and Carriquiry (2019) urge that researchers prominently display the probability the hypothesis is true or a probability distribution of an effect size, or provide sufficient information for future researchers and policy makers to compute it. The authors further describe why such a probability is necessary for decision making, how it could be estimated by using historical rates of reproduction of findings, and how this same process can be part of continuous "quality control" for science.

Being thoughtful in our approach to research will lead us to **be open** in our design, conduct, and presentation of it as well.

3.3. Be Open

We envision **openness** as embracing certain positive practices in the development and presentation of research work.

3.3.1. Openness to Transparency and to the Role of Expert Judgment

First, we repeat oft-repeated advice: **Be open** to "open science" practices. Calin-Jageman and Cumming (2019), Locascio (2019), and others in this special issue urge adherence to practices such as public pre-registration of methods, transparency and completeness in reporting, shared data and code, and even pre-registered ("results-blind") review. Completeness in reporting, for example, requires not only describing all analyses performed but also presenting all findings obtained, without regard to statistical significance or any such criterion.

Openness also includes understanding and accepting the role of expert judgment, which enters the practice of statistical inference and decision-making in numerous ways (O'Hagan 2019). "Indeed, there is essentially no aspect of scientific investigation in which judgment is not required," O'Hagan observes. "Judgment is necessarily subjective, but should be made as carefully, as objectively, and as scientifically as possible."

Subjectivity is involved in any statistical analysis, Bayesian or frequentist. Gelman and Hennig (2017) observe, "Personal decision making cannot be avoided in statistical data analysis and, for want of approaches to justify such decisions, the pursuit of objectivity degenerates easily to a pursuit to merely *appear* objective." One might say that subjectivity is not a problem; it is part of the solution.

Acknowledging this, Brownstein et al. (2019) point out that expert judgment and knowledge are required in all stages of the scientific method. They examine the roles of expert judgment throughout the scientific process, especially regarding the integration of statistical and content expertise. "All researchers, irrespective of their philosophy or practice, use expert judgment in developing models and interpreting results," say Brownstein et al. "We must accept that there is subjectivity in every stage of scientific inquiry, but objectivity is nevertheless the fundamental goal. Therefore, we should base judgments on evidence and careful reasoning, and seek wherever possible to eliminate potential sources of bias."

How does one rigorously elicit expert knowledge and judgment in an effective, unbiased, and transparent way? O'Hagan (2019) addresses this, discussing protocols to elicit expert knowledge in an unbiased and as scientifically sound was as possible. It is also important for such elicited knowledge to be examined critically, comparing it to actual study results being an important diagnostic step.

3.3.2. Openness in Communication

Be open in your reporting. Report p-values as continuous, descriptive statistics, as we explain in Section 2. We realize that this leaves researchers without their familiar bright line anchors. Yet if we were to propose a universal template for presenting and interpreting continuous p-values we would violate our own principles! Rather, we believe that the thoughtful use and interpretation of p-values will never adhere to a rigid rulebook, and will instead inevitably vary from study to study. Despite these caveats, we can offer recommendations for sound practices, as described below.

In all instances, regardless of the value taken by p or any other statistic, consider what McShane et al. (2019) call the "currently subordinate factors"—the factors that should no longer be subordinate to "$p < 0.05$." These include relevant prior evidence, plausibility of mechanism, study design and data quality, and the real-world costs and benefits that determine what effects are scientifically important. The scientific context of your study matters, they say, and this should guide your interpretation.

When using p-values, remember not only Principle 5 of the ASA statement: "A p-value…does not measure the size of an effect or the importance of a result" but also Principle 6: "By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis." Despite these limitations, if you present p-values, do so for more than one hypothesized value of your variable of interest (Fraser 2019; Greenland 2019), such as 0 and at least one plausible, relevant alternative, such as the minimum practically important effect size (which should be determined before analyzing the data).

Betensky (2019) also reminds us to interpret the p-value in the context of sample size and meaningful effect size.

Instead of p, you might consider presenting the s-value (Greenland 2019), which is described in Section 3.2. As noted in Section 3.1, you might present a confidence interval. Sound practices in the interpretation of confidence intervals include (1) discussing both the upper and lower limits and whether they have different practical implications, (2) paying no particular attention to whether the interval includes the null value, and (3) remembering that an interval is itself an estimate subject to error and generally provides only a rough indication of uncertainty given that all of the assumptions used to create it are correct and, thus, for example, does not "rule out" values outside the interval. Amrhein, Trafimow, and Greenland (2019) suggest that interval estimates be interpreted as "compatibility" intervals rather than as "confidence" intervals, showing the values that are most compatible with the data, under the model used to compute the interval. They argue that such an interpretation and the practices outlined here can help guard against overconfidence.

It is worth noting that Tong (2019) disagrees with using p-values as descriptive statistics. "Divorced from the probability claims attached to such quantities (confidence levels, nominal Type I errors, and so on), there is no longer any reason to privilege such quantities over descriptive statistics that more directly characterize the data at hand." He further states, "Methods with alleged generality, such as the p-value or Bayes factor, should be avoided in favor of discipline- and problem-specific solutions that can be designed to be fit for purpose."

Failing to **be open** in reporting leads to publication bias. Ioannidis (2019) notes the high level of selection bias prevalent in biomedical journals. He defines "selection" as "the collection of choices that lead from the planning of a study to the reporting of p-values." As an illustration of one form of selection bias, Ioannidis compared "the set of p-values reported in the full text of an article with the set of p-values reported in the abstract." The main finding, he says, "was that p-values chosen for the abstract tended to show greater significance than those reported in the text, and that the gradient was more pronounced in some types of journals and types of designs." Ioannidis notes, however, that selection bias "can be present regardless of the approach to inference used." He argues that in the long run, "the only direct protection must come from standards for reproducible research."

To **be open**, remember that one study is rarely enough. The words "a groundbreaking new study" might be loved by news writers but must be resisted by researchers. Breaking ground is only the first step in building a house. It will be suitable for habitation only after much more hard work.

Be open by providing sufficient information so that other researchers can execute meaningful alternative analyses. van Dongen et al. (2019) provide an illustrative example of such alternative analyses by different groups attacking the same problem.

Being open goes hand in hand with **being modest.**

3.4. Be Modest

Researchers of any ilk may rarely advertise their personal modesty. Yet the most successful ones cultivate a practice of **being modest** throughout their research, by understanding and clearly expressing the limitations of their work.

Being modest requires a reality check (Amrhein, Trafimow, and Greenland 2019). "A core problem," they observe, "is that both scientists and the public confound statistics with reality. But statistical inference is a thought experiment, describing the predictive performance of models about reality. Of necessity, these models are extremely simplified relative to the complexities of actual study conduct and of the reality being studied. Statistical results must eventually mislead us when they are used and communicated as if they present this complex reality, rather than a model for it. This is not a problem of our statistical methods. It is a problem of interpretation and communication of results."

Be modest in recognizing there is not a "true statistical model" underlying every problem, which is why it is wise to **thoughtfully** consider many possible models (Lavine 2019). Rougier (2019) calls on researchers to "recognize that behind every choice of null distribution and test statistic, there lurks

a plausible family of alternative hypotheses, which can provide more insight into the null distribution."

p-values, confidence intervals, and other statistical measures are all uncertain. Treating them otherwise is immodest overconfidence.

Remember that statistical tools have their limitations. Rose and McGuire (2019) show how use of stepwise regression in health care settings can lead to policies that are unfair.

Remember also that the amount of evidence for or against a hypothesis provided by *p*-values near the ubiquitous $p < 0.05$ threshold (Johnson 2019) is usually much less than you think (Benjamin and Berger 2019; Colquhoun 2019; Greenland 2019).

Be modest about the role of statistical inference in scientific inference. "Scientific inference is a far broader concept than statistical inference," says Hubbard, Haig, and Parsa (2019). "A major focus of scientific inference can be viewed as the pursuit of *significant sameness*, meaning replicable and empirically generalizable results among phenomena. Regrettably, the obsession with users of statistical inference to report *significant differences* in data sets actively thwarts cumulative knowledge development."

The nexus of **openness** and **modesty** is to report everything while at the same time not concluding anything from a single study with unwarranted certainty. Because of the strong desire to inform and be informed, there is a relentless demand to state results with certainty. Again, **accept uncertainty** and embrace variation in associations and effects, because they are always there, like it or not. Understand that expressions of uncertainty are themselves uncertain. Accept that one study is rarely definitive, so encourage, sponsor, conduct, and publish replication studies. Then, use meta-analysis, evidence reviews, and Bayesian methods to synthesize evidence across studies.

Resist the urge to overreach in the generalizability of claims. Watch out for pressure to embellish the abstract or the press release. If the study's limitations are expressed in the paper but not in the abstract, they may never be read.

Be modest by encouraging others to reproduce your work. Of course, for it to be reproduced readily, you will necessarily have been **thoughtful** in conducting the research and **open** in presenting it.

Hubbard and Carriquiry (see their "do list" in Section 7) suggest encouraging reproduction of research by giving "a byline status for researchers who reproduce studies." They would like to see digital versions of papers dynamically updated to display "Reproduced by…." below original research authors' names or "not yet reproduced" until it is reproduced.

Indeed, when it comes to reproducibility, Amrhein, Trafimow, and Greenland (2019) demand that we **be modest** in our expectations. "An important role for statistics in research is the summary and accumulation of information," they say. "If replications do not find the same results, this is not necessarily a crisis, but is part of a natural process by which science evolves. The goal of scientific methodology should be to direct this evolution toward ever more accurate descriptions of the world and how it works, not toward ever more publication of inferences, conclusions, or decisions."

Referring to replication studies in psychology, McShane et al. (2019) recommend that future large-scale replication projects "should follow the 'one phenomenon, many studies' approach of the Many Labs project and Registered Replication Reports rather than the 'many phenomena, one study' approach of the Open Science Collaboration project. In doing so, they should systematically vary method factors across the laboratories involved in the project." This approach helps achieve the goals of Amrhein, Trafimow, and Greenland (2019) by increasing understanding of why and when results replicate or fail to do so, yielding more accurate descriptions of the world and how it works. It also speaks to significant sameness versus significant difference a la Hubbard, Haig, and Parsa (2019).

Kennedy-Shaffer's (2019) historical perspective on statistical significance reminds us to **be modest**, by prompting us to recall how the current state of affairs in *p*-values has come to be.

Finally, **be modest** by recognizing that different readers may have very different stakes on the results of your analysis, which means you should try to take the role of a neutral judge rather than an advocate for any hypothesis. This can be done, for example, by pairing every null *p*-value with a *p*-value testing an equally reasonable alternative, and by discussing the endpoints of every interval estimate (not only whether it contains the null).

Accept that both scientific inference and statistical inference are hard, and understand that no knowledge will be efficiently advanced using simplistic, mechanical rules and procedures. Accept also that pure objectivity is an unattainable goal—no matter how laudable—and that both subjectivity and expert judgment are intrinsic to the conduct of science and statistics. Accept that there will always be uncertainty, and be **t**houghtful, **o**pen, and **m**odest, ATOM.

And to push this acronym further, we argue in the next section that institutional change is needed, so we put forward that change is needed at the ATOMIC level. Let's go.

4. Editorial, Educational and Other Institutional Practices Will Have to Change

Institutional reform is necessary for moving beyond statistical significance in any context—whether journals, education, academic incentive systems, or others. Several papers in this special issue focus on reform.

Goodman (2019) notes considerable social change is needed in academic institutions, in journals, and among funding and regulatory agencies. He suggests (see Section 7) partnering "with science reform movements and reformers within disciplines, journals, funding agencies and regulators to promote and reward 'reproducible' science and diminish the impact of statistical significance on publication, funding and promotion." Similarly, Colquhoun (2019) says, "In the end, the only way to solve the problem of reproducibility is to do more replication and to reduce the incentives that are imposed on scientists to produce unreliable work. The publish-or-perish culture has damaged science, as has the judgment of their work by silly metrics."

Trafimow (2019), who added energy to the discussion of *p*-values a few years ago by banning them from the journal he edits (Fricker et al. 2019), suggests five "nonobvious changes" to editorial practice. These suggestions, which demand reevaluating traditional practices in editorial policy, will not be trivial to implement but would result in massive change in some journals.

Locascio (2017, 2019) suggests that evaluation of manuscripts for publication should be "results-blind." That is, manuscripts should be assessed for suitability for publication based on the substantive importance of the research without regard to their reported results. Kmetz (2019) supports this approach as well and says that it would be a huge benefit for reviewers, "freeing [them] from their often thankless present jobs and instead allowing them to review research designs for their potential to provide useful knowledge." (See also "registered reports" from the Center for Open Science (*https://cos.io/rr/?_ga=2.184185454.979594832.1547755516-1193527346.1457026171*) and "registered replication reports" from the Association for Psychological Science (*https://www.psychologicalscience.org/publications/replication*) in relation to this concept.)

Amrhein, Trafimow, and Greenland (2019) ask if results-blind publishing means that anything goes, and then answer affirmatively: "Everything should be published in some form if whatever we measured made sense *before we obtained the data* because it was connected in a potentially useful way to some research question." Journal editors, they say, "should be proud about [their] exhaustive methods sections" and base their decisions about the suitability of a study for publication "on the quality of its materials and methods rather than on results and conclusions; the quality of the presentation of the latter is only judged after it is determined that the study is valuable based on its materials and methods."

A "variation on this theme is *pre-registered replication*, where a *replication* study, rather than the original study, is subject to strict pre-registration (e.g., Gelman 2015)," says Tong (2019). "A broader vision of this idea (Mogil and Macleod 2017) is to carry out a whole series of exploratory experiments *without* any formal statistical inference, and summarize the results by descriptive statistics (including graphics) or even just disclosure of the raw data. When results from this series of experiments converge to a single working hypothesis, it can *then* be subjected to a pre-registered, randomized and blinded, appropriately powered confirmatory experiment, carried out by another laboratory, in which valid statistical inference may be made."

Hurlbert, Levine, and Utts (2019) urge abandoning the use of "statistically significant" in all its forms and encourage journals to provide instructions to authors along these lines: "There is now wide agreement among many statisticians who have studied the issue that for reporting of statistical tests yielding *p*-values it is illogical and inappropriate to dichotomize the *p*-scale and describe results as 'significant' and 'nonsignificant.' Authors are strongly discouraged from continuing this never justified practice that originated from confusions in the early history of modern statistics."

Hurlbert, Levine, and Utts (2019) also urge that the *ASA Statement on P-Values and Statistical Significance* "be sent to the editor-in-chief of every journal in the natural, behavioral and social sciences for forwarding to their respective editorial boards and stables of manuscript reviewers. That would be a good way to quickly improve statistical understanding and practice." Kmetz (2019) suggests referring to the ASA statement whenever submitting a paper or revision to any editor, peer reviewer, or prospective reader. Hurlbert et al. encourage a "community grassroots effort" to encourage change in journal procedures.

Campbell and Gustafson (2019) propose a statistical model for evaluating publication policies in terms of weighing novelty of studies (and the likelihood of those studies subsequently being found false) against pre-specified study power. They observe that "no publication policy will be perfect. Science is inherently challenging and we must always be willing to accept that a certain proportion of research is potentially false."

Statistics education will require major changes at all levels to move to a post "$p < 0.05$" world. Two papers in this special issue make a specific start in that direction (Maurer et al. 2019; Steel, Liermann, and Guiltorp 2019), but we hope that volumes will be written on this topic in other venues. We are excited that, with support from the ASA, the US Conference on Teaching Statistics (USCOTS) will focus its 2019 meeting on teaching inference.

The change that needs to happen demands change to editorial practice, to the teaching of statistics at every level where inference is taught, and to much more. However...

5. It Is Going to Take Work, and It Is Going to Take Time

If it were easy, it would have already been done, because as we have noted, this is nowhere near the first time the alarm has been sounded.

Why is eliminating the use of *p*-values as a truth arbiter so hard? "The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them," says Goodman (2019). "It's the same reason we can use money. When everyone believes in something's value, we can use it for real things; money for food, and *p*-values for knowledge claims, publication, funding, and promotion. It doesn't matter if the *p*-value doesn't mean what people think it means; it becomes valuable because of what it buys."

Goodman observes that statisticians alone cannot address the problem, and that "any approach involving only statisticians will not succeed." He calls on statisticians to ally themselves "both with scientists in other fields and with broader based, multidisciplinary scientific reform movements. What statisticians can do within our own discipline is important, but to effectively disseminate or implement virtually any method or policy, we need partners."

"The loci of influence," Goodman says, "include journals, scientific lay and professional media (including social media), research funders, healthcare payors, technology assessors, regulators, academic institutions, the private sector, and professional societies. They also can include policy or informational entities like the National Academies...as well as various other science advisory bodies across the government. Increasingly, they are also including non-traditional science reform organizations comprised both of scientists and of the science literate lay public...and a broad base of health or science advocacy groups..."

It is no wonder, then, that the problem has persisted for so long. And persist it has! Hubbard (2019) looked at citation-count data on twenty-five articles and books severely critical of the effect of null hypothesis significance testing (NHST) on good science. Though issues were well known, Hubbard says, this did nothing to stem NHST usage over time.

Greenland (personal communication, January 25, 2019) notes that cognitive biases and perverse incentives to offer firm conclusions where none are warranted can warp the use of any method. "The core human and systemic problems are not addressed by shifting blame to p-values and pushing alternatives as magic cures—especially alternatives that have been subject to little or no comparative evaluation in either classrooms or practice." Greenland said. "What we need now is to move beyond debating only our methods and their interpretations, to concrete proposals for elimination of systemic problems such as pressure to produce noteworthy findings rather than to produce reliable studies and analyses. Review and provisional acceptance of reports before their results are given to the journal (Locascio 2019) is one way to address that pressure, but more ideas are needed since review of promotions and funding applications cannot be so blinded. The challenges of how to deal with human biases and incentives may be the most difficult we must face." Supporting this view is McShane and Gal's (2016, 2017) empirical demonstration of cognitive dichotomization errors among biomedical and social science researchers—and even among statisticians.

Challenges for editors and reviewers are many. Here's an example: Fricker et al. (2019) observed that when p-values were suspended from the journal *Basic and Applied Social Psychology* authors tended to overstate conclusions.

With all the challenges, how do we get from here to there, from a "$p < 0.05$" world to a post "$p < 0.05$" world?

Matthews (2019) notes that "Any proposal encouraging changes in inferential practice must accept the ubiquity of NHST.... Pragmatism suggests, therefore, that the best hope of achieving a change in practice lies in offering inferential tools that can be used alongside the concepts of NHST, adding value to them while mitigating their most egregious features."

Benjamin and Berger (2019) propose three practices to help researchers during the transition away from use of statistical significance. "...[O]ur goal is to suggest minimal changes that would require little effort for the scientific community to implement," they say. "Motivating this goal are our hope that easy (but impactful) changes might be adopted and our worry that more complicated changes could be resisted simply because they are perceived to be too difficult for routine implementation."

Yet there is also concern that progress will stop after a small step or two. Even some proponents of small steps are clear that those small steps still carry us far short of the destination.

For example, Matthews (2019) says that his proposed methodology "is not a panacea for the inferential ills of the research community." But that doesn't make it useless. It may "encourage researchers to move beyond NHST and explore the statistical armamentarium now available to answer the central question of research: what does our study tell us?" he says. It "provides a bridge between the dominant but flawed NHST paradigm and the less familiar but more informative methods of Bayesian estimation."

Likewise, Benjamin and Berger (2019) observe, "In research communities that are deeply attached to reliance on '$p < 0.05$,' our recommendations will serve as initial steps away from this attachment. We emphasize that our recommendations are intended merely as initial, temporary steps and that many further steps will need to be taken to reach the ultimate destination: a holistic interpretation of statistical evidence that fully conforms to the principles laid out in the ASA Statement..."

Yet, like the authors of this editorial, not all authors in this special issue support gradual approaches with transitional methods.

Some (e.g., Amrhein, Trafimow, and Greenland 2019; Hurlbert, Levine, and Utts 2019; McShane et al. 2019) prefer to rip off the bandage and abandon use of statistical significance altogether. In short, no more dichotomizing p-values into categories of "significance." Notably, these authors do not suggest banning the use of p-values, but rather suggest using them descriptively, treating them as continuous, and assessing their weight or import with nuanced thinking, clear language, and full understanding of their properties.

So even when there is agreement on the destination, there is disagreement about what road to take. The questions around reform need consideration and debate. It might turn out that different fields take different roads.

The catalyst for change may well come from those people who fund, use, or depend on scientific research, say Calin-Jageman and Cumming (2019). They believe this change has not yet happened to the desired level because of "the cognitive opacity of the NHST approach: the counter-intuitive p-value (it's good when it is small), the mysterious null hypothesis (you want it to be false), and the eminently confusable Type I and Type II errors."

Reviewers of this editorial asked, as some readers of it will, is a p-value threshold ever okay to use? We asked some of the authors of articles in the special issue that question as well. Authors identified four general instances. Some allowed that, while p-value thresholds should not be used for inference, they might still be useful for applications such as industrial quality control, in which a highly automated decision rule is needed and the costs of erroneous decisions can be carefully weighed when specifying the threshold. Other authors suggested that such dichotomized use of p-values was acceptable in model-fitting and variable selection strategies, again as automated tools, this time for sorting through large numbers of potential models or variables. Still others pointed out that p-values with very low thresholds are used in fields such as physics, genomics, and imaging as a filter for massive numbers of tests. The fourth instance can be described as "confirmatory setting[s] where the study design and statistical analysis plan are specified prior to data collection, and then adhered to during and after it" (Tong 2019). Tong argues these are the only proper settings for formal statistical inference. And Wellek (2017) says at present it is essential in these settings. "[B]inary decision making is indispensable in medicine and related fields," he says. "[A] radical rejection of the classical principles of statistical inference...is of virtually no help as long as no conclusively substantiated alternative can be offered."

Eliminating the declaration of "statistical significance" based on $p < 0.05$ or other arbitrary thresholds will be easier in some venues than others. Most journals, if they are willing, could fairly rapidly implement editorial policies to effect these changes. Suggestions for how to do that are in this special issue of *The American Statistician*. However, regulatory agencies might require longer timelines for making changes. The U.S. Food and

Drug Administration (FDA), for example, has long established drug review procedures that involve comparing p-values to significance thresholds for Phase III drug trials. Many factors demand consideration, not the least of which is how to avoid turning every drug decision into a court battle. Goodman (2019) cautions that, even as we seek change, "we must respect the reason why the statistical procedures are there in the first place." Perhaps the ASA could convene a panel of experts, internal and external to FDA, to provide a workable new paradigm. (See Ruberg et al. 2019, who argue for a Bayesian approach that employs data from other trials as a "prior" for Phase 3 trials.)

Change is needed. Change has been needed for decades. Change has been called for by others for quite a while. So...

6. Why Will Change Finally Happen Now?

In 1991, a confluence of weather events created a monster storm that came to be known as "the perfect storm," entering popular culture through a book (Junger 1997) and a 2000 movie starring George Clooney. Concerns about reproducible science, falling public confidence in science, and the initial impact of the ASA statement in heightening awareness of long-known problems created a perfect storm, in this case, a good storm of motivation to make lasting change. Indeed, such change was the intent of the ASA statement, and we expect this special issue of TAS will inject enough additional energy to the storm to make its impact widely felt.

We are not alone in this view. "60+ years of incisive criticism has not yet dethroned NHST as the dominant approach to inference in many fields of science," note Calin-Jageman and Cumming (2019). "Momentum, though, seems to finally be on the side of reform."

Goodman (2019) agrees: "The initial slow speed of progress should not be discouraging; that is how all broad-based social movements move forward and we should be playing the long game. But the ball is rolling downhill, the current generation is inspired and impatient to carry this forward."

So, let's do it. Let's move beyond "statistically significant," even if upheaval and disruption are inevitable for the time being. It's worth it. In a world beyond "$p < 0.05$," by breaking free from the bonds of statistical significance, statistics in science and policy will become more significant than ever.

7. Authors' Suggestions

The editors of this special TAS issue on statistical inference asked all the contact authors to help us summarize the guidance they provided in their papers by providing us a short list of do's. We asked them to be specific but concise and to be active—start each with a verb. Here is the complete list of the authors' responses, ordered as the papers appear in this special issue.

7.1. Getting to a Post "$p < 0.05$" Era

Ioannidis, J., What Have We (Not) Learnt From Millions of Scientific Papers With p-Values?

1. Do not use p-values, unless you have clearly thought about the need to use them and they still seem the best choice.
2. Do not favor "statistically significant" results.
3. Do be highly skeptical about "statistically significant" results at the 0.05 level.

Goodman, S., Why Is Getting Rid of p-Values So Hard? Musings on Science and Statistics

1. Partner with science reform movements and reformers within disciplines, journals, funding agencies and regulators to promote and reward reproducible science and diminish the impact of statistical significance on publication, funding and promotion.
2. Speak to and write for the multifarious array of scientific disciplines, showing how statistical uncertainty and reasoning can be conveyed in non-"bright-line" ways both with conventional and alternative approaches. This should be done not just in didactic articles, but also in original or reanalyzed research, to demonstrate that it is publishable.
3. Promote, teach and conduct meta-research within many individual scientific disciplines to demonstrate the adverse effects in each of over-reliance on and misinterpretation of p-values and significance verdicts in individual studies and the benefits of emphasizing estimation and cumulative evidence.
4. Require reporting a quantitative measure of certainty—a "confidence index"—that an observed relationship, or claim, is true. Change analysis goal from achieving significance to appropriately estimating this confidence.
5. Develop and share teaching materials, software, and published case examples to help with all of the do's above, and to spread progress in one discipline to others.

Hubbard, R., Will the ASA's Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary

This list applies to the ASA and to the professional statistics community more generally.

1. Specify, where/if possible, those situations in which the p-value plays a clearly valuable role in data analysis and interpretation.
2. Contemplate issuing a statement abandoning the use of p-values in null hypothesis significance testing.

Kmetz, J., Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of p-Values

1. Refer to the ASA statement on p-values whenever submitting a paper or revision to any editor, peer reviewer, or prospective reader. Many in the field do not know of this statement, and having the support of a prestigious organization when authoring any research document will help stop corrupt research from becoming even more dominant than it is.
2. Train graduate students and future researchers by having them reanalyze published studies and post their findings to appropriate websites or weblogs. This practice will benefit not only the students, but will benefit the professions, by increasing the amount of replicated (or nonreplicated) research available and readily accessible, and as well as reformer organizations that support replication.
3. Join one or more of the reformer organizations formed or forming in many research fields, and support and publicize their efforts to improve the quality of research practices.

4. Challenge editors and reviewers when they assert that incorrect practices and interpretations of research, consistent with existing null hypothesis significance testing and beliefs regarding *p*-values, should be followed in papers submitted to their journals. Point out that new submissions have been prepared to be consistent with the ASA statement on *p*-values.
5. Promote emphasis on research quality rather than research quantity in universities and other institutions where professional advancement depends heavily on research "productivity," by following the practices recommended in this special journal edition. This recommendation will fall most heavily on those who have already achieved success in their fields, perhaps by following an approach quite different from that which led to their success; whatever the merits of that approach may have been, one objectionable outcome of it has been the production of voluminous corrupt research and creation of an environment that promotes and protects it. We must do better.

Hubbard, D., and Carriquiry, A., Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness and Relevance

1. Compute and prominently display the probability the hypothesis is true (or a probability distribution of an effect size) or provide sufficient information for future researchers and policy makers to compute it.
2. Promote publicly displayed quality control metrics within your field—in particular, support tracking of reproduction studies and computing the "level 1" and even "level 2" priors as required for #1 above.
3. Promote a byline status for researchers who reproduce studies: Digital versions are dynamically updated to display "Reproduced by…." below original research authors' names or "Not yet reproduced" until it is reproduced.

Brownstein, N., Louis, T., O'Hagan, A., and Pendergast, J., The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making

1. Staff the study team with members who have the necessary knowledge, skills and experience—statistically, scientifically, and otherwise.
2. Include key members of the research team, including statisticians, in all scientific and administrative meetings.
3. Understand that subjective judgments are needed in all stages of a study.
4. Make all judgments as carefully and rigorously as possible and document each decision and rationale for transparency and reproducibility.
5. Use protocol-guided elicitation of judgments.
6. Statisticians specifically should:
   - Refine oral and written communication skills.
   - Understand their multiple roles and obligations as collaborators.
   - Take an active leadership role as a member of the scientific team; contribute throughout all phases of the study.
   - Co-own the subject matter—understand a sufficient amount about the relevant science/policy to meld statistical and subject-area expertise.
   - Promote the expectation that your collaborators co-own statistical issues.
   - Write a statistical analysis plan for all analyses and track any changes to that plan over time.
   - Promote co-responsibility for data quality, security, and documentation.
   - Reduce unplanned and uncontrolled modeling/testing (HARK-ing, *p*-hacking); document all analyses.

O'Hagan, A., Expert Knowledge Elicitation: Subjective but Scientific

1. Elicit expert knowledge when data relating to a parameter of interest is weak, ambiguous or indirect.
2. Use a well-designed protocol, such as SHELF, to ensure expert knowledge is elicited in as scientific and unbiased a way as possible.

Kennedy-Shaffer, L., Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing

1. Ensure that inference methods match intuitive understandings of statistical reasoning.
2. Reduce the computational burden for nonstatisticians using statistical methods.
3. Consider changing conditions of statistical and scientific inference in developing statistical methods.
4. Address uncertainty quantitatively and in ways that reward increased precision.

Hubbard, R., Haig, B. D., and Parsa, R. A., The Limited Role of Formal Statistical Inference in Scientific Inference

1. Teach readers that although deemed equivalent in the social, management, and biomedical sciences, formal methods of statistical inference and scientific inference are very different animals.
2. Show these readers that formal methods of statistical inference play only a restricted role in scientific inference.
3. Instruct researchers to pursue significant *sameness* (i.e., replicable and empirically generalizable results) rather than significant *differences* in results.
4. Demonstrate how the pursuit of significant differences actively impedes cumulative knowledge development.

McShane, B., Tackett, J., Böckenholt, U., and Gelman, A., Large Scale Replication Projects in Contemporary Psychological Research

1. When planning a replication study of a given psychological phenomenon, bear in mind that replication is complicated in psychological research because studies can never be direct or exact replications of one another, and thus heterogeneity—effect sizes that vary from one study of the phenomenon to the next—cannot be avoided.
2. Future large scale replication projects should follow the "one phenomenon, many studies" approach of the Many Labs project and Registered Replication Reports rather than the

"many phenomena, one study" approach of the Open Science Collaboration project. In doing so, they should systematically vary method factors across the laboratories involved in the project.
3. Researchers analyzing the data resulting from large scale replication projects should do so via a hierarchical (or multilevel) model fit to the totality of the individual-level observations. In doing so, all theoretical moderators should be modeled via covariates while all other potential moderators—that is, method factors—should induce variation (i.e., heterogeneity).
4. Assessments of replicability should not depend solely on estimates of effects, or worse, significance tests based on them. Heterogeneity must also be an important consideration in assessing replicability.

7.2. Interpreting and Using p

Greenland, S., Valid p-Values Behave Exactly as They Should: Some Misleading Criticisms of p-Values and Their Resolution With s-Values

1. Replace any statements about statistical significance of a result with the p-value from the test, and present the p-value as an equality, not an inequality. For example, if $p = 0.03$ then "...was statistically significant" would be replaced by "...had $p = 0.03$," and "$p < 0.05$" would be replaced by "$p = 0.03$." (An exception: If p is so small that the accuracy becomes very poor then an inequality reflecting that limit is appropriate; e.g., depending on the sample size, p-values from normal or χ^2 approximations to discrete data often lack even 1-digit accuracy when $p < 0.0001$.) In parallel, if $p = 0.25$ then "...was not statistically significant" would be replaced by "...had $p = 0.25$," and "$p > 0.05$" would be replaced by "$p = 0.25$."
2. Present p-values for more than one possibility when testing a targeted parameter. For example, if you discuss the p-value from a test of a null hypothesis, also discuss alongside this null p-value another p-value for a plausible alternative parameter possibility (ideally the one used to calculate power in the study proposal). As another example: if you do an equivalence test, present the p-values for both the lower and upper bounds of the equivalence interval (which are used for equivalence tests based on two one-sided tests).
3. Show confidence intervals for targeted study parameters, but also supplement them with p-values for testing relevant hypotheses (e.g., the p-values for both the null and the alternative hypotheses used for the study design or proposal, as in #2). Confidence intervals only show clearly what is in or out of the interval (i.e., a 95% interval only shows clearly what has $p > 0.05$ or $p \leq 0.05$), but more detail is often desirable for key hypotheses under contention.
4. Compare groups and studies directly by showing p-values and interval estimates for their differences, not by comparing p-values or interval estimates from the two groups or studies. For example, seeing $p = 0.03$ in males and $p = 0.12$ in females does ***not*** mean that different associations were seen in males and females; instead, one needs a p-value and confidence interval for the difference in the sex-specific associations to examine the between-sex difference. Similarly, if an early study reported a confidence interval which excluded the null and then a subsequent study reported a confidence interval which included the null, that does not mean the studies gave conflicting results or that the second study failed to replicate the first study; instead, one needs a p-value and confidence interval for the difference in the study-specific associations to examine the between-study difference. In all cases, differences-between-differences must be analyzed directly by statistics for that purpose.
5. Supplement a focal p-value p with its Shannon information transform (s-value or surprisal) $s = -\log_2(p)$. This measures the amount of information supplied by the test against the tested hypothesis (or model): Rounded off, the s-value s shows the number of heads in a row one would need to see when tossing a coin to get the same amount of information against the tosses being "fair" (independent with "heads" probability of 1/2) instead of being loaded for heads. For example, if $p = 0.03$, this represents $-\log_2(0.03) = 5$ bits of information against the hypothesis (like getting 5 heads in a trial of "fairness" with 5 coin tosses); and if $p = 0.25$, this represents only $-\log_2(0.25) = 2$ bits of information against the hypothesis (like getting 2 heads in a trial of "fairness" with only 2 coin tosses).

Betensky, R., The p-Value Requires Context, Not a Threshold

1. Interpret the p-value in light of its context of sample size and meaningful effect size.
2. Incorporate the sample size and meaningful effect size into a decision to reject the null hypothesis.

Anderson, A., Assessing Statistical Results: Magnitude, Precision and Model Uncertainty

1. Evaluate the importance of statistical results based on their practical implications.
2. Evaluate the strength of empirical evidence based on the precision of the estimates and the plausibility of the modeling choices.
3. Seek out subject matter expertise when evaluating the importance and the strength of empirical evidence.

Heck, P., and Krueger, J., Putting the p-Value in Its Place

1. Use the p-value as a heuristic, that is, as the base for a tentative inference regarding the presence or absence of evidence against the tested hypothesis.
2. Supplement the p-value with other, conceptually distinct methods and practices, such as effect size estimates, likelihood ratios, or graphical representations.
3. Strive to embed statistical hypothesis testing within strong *a priori* theory and a context of relevant prior empirical evidence.

Johnson, V., Evidence From Marginally Significant t-Statistics

1. Be transparent in the number of outcome variables that were analyzed.
2. Report the number (and values) of all test statistics that were calculated.
3. Provide access to protocols for studies involving human or animal subjects.

4. Clearly describe data values that were excluded from analysis and the justification for doing so.
5. Provide sufficient details on experimental design so that other researchers can replicate the experiment.
6. Describe only p-values less than 0.005 as being "statistically significant."

Fraser, D., The p-Value Function and Statistical Inference

1. Determine a primary variable for assessing the hypothesis at issue.
2. Calculate its well defined distribution function, respecting continuity.
3. Substitute the observed data value to obtain the "p-value function."
4. Extract the available well defined confidence bounds, confidence intervals, and median estimate.
5. Know that you don't have an intellectual basis for decisions.

Rougier, J., p-Values, Bayes Factors, and Sufficiency

1. Recognize that behind every choice of null distribution and test statistic, there lurks a plausible family of alternative hypotheses, which can provide more insight into the null distribution.

Rose, S., and McGuire, T., Limitations of p-Values and R-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment

1. Formulate a clear objective for variable inclusion in regression procedures.
2. Assess all relevant evaluation metrics.
3. Incorporate algorithmic fairness considerations.

7.3. Supplementing or Replacing p

Blume, J., Greevy, R., Welty, V., Smith, J., and DuPont, W., An Introduction to Second Generation p-Values

1. Construct a composite null hypothesis by specifying the range of effects that are not scientifically meaningful (do this before looking at the data). Why: Eliminating the conflict between scientific significance and statistical significance has numerous statistical and scientific benefits.
2. Replace classical p-values with second-generation p-values (SGPV). Why: SGPVs accommodate composite null hypotheses and encourage the proper communication of findings.
3. Interpret the SGPV as a high-level summary of what the data say. Why: Science needs a simple indicator of when the data support only meaningful effects (SGPV = 0), when the data support only trivially null effects (SGPV = 1), or when the data are inconclusive ($0 < \text{SGPV} < 1$).
4. Report an interval estimate of effect size (confidence interval, support interval, or credible interval) and note its proximity to the composite null hypothesis. Why: This is a more detailed description of study findings.
5. Consider reporting false discovery rates with SGPVs of 0 or 1. Why: FDRs gauge the chance that an inference is incorrect under assumptions about the data generating process and prior knowledge.

Goodman, W., Spruill, S., and Komaroff, E., A Proposed Hybrid Effect Size Plus p-Value Criterion: Empirical Evidence Supporting Its Use

1. Determine how far the true parameter's value would have to be, in your research context, from exactly equaling the conventional, point null hypothesis to consider that the distance is meaningfully large or practically significant.
2. Combine the conventional p-value criterion with a minimum effect size criterion to generate a two-criteria inference-indicator signal, which provides heuristic, but nondefinitive evidence, for inferring the parameter's true location.
3. Document the intended criteria for your inference procedures, such as a p-value cut-point and a minimum practically significant effect size, prior to undertaking the procedure.
4. Ensure that you use the appropriate inference method for the data that are obtainable and for the inference that is intended.
5. Acknowledge that every study is fraught with limitations from unknowns regarding true data distributions and other conditions that one's method assumes.

Benjamin, D., and Berger, J., Three Recommendations for Improving the Use of p-Values

1. Replace the 0.05 "statistical significance" threshold for claims of novel discoveries with a 0.005 threshold and refer to p-values between 0.05 and 0.005 as "suggestive."
2. Report the data-based odds of the alternative hypothesis to the null hypothesis. If the data-based odds cannot be calculated, then use the p-value to report an upper bound on the data-based odds: $1/(-ep \ln p)$.
3. Report your prior odds and posterior odds (prior odds * data-based odds) of the alternative hypothesis to the null hypothesis. If the data-based odds cannot be calculated, then use your prior odds and the p-value to report an upper bound on your posterior odds: (prior odds) * $(1/(-ep \ln p))$.

Colquhoun, D., The False Positive Risk: A Proposal Concerning What to Do About p-Values

1. Continue to provide p-values and confidence intervals. Although widely misinterpreted, people know how to calculate them and they aren't entirely useless. Just don't ever use the terms "statistically significant" or "nonsignificant."
2. Provide in addition an indication of false positive risk (FPR). This is the probability that the claim of a real effect on the basis of the p-value is in fact false. The FPR (not the p-value) is the probability that your result occurred by chance. For example, the fact that, under plausible assumptions, observation of a p-value close to 0.05 corresponds to an FPR of at least 0.2–0.3 shows clearly the weakness of the conventional criterion for "statistical significance."
3. Alternatively, specify the prior probability of there being a real effect that one would need to be able to justify in order to achieve an FPR of, say, 0.05.

Notes:

There are many ways to calculate the FPR. One, based on a point null and simple alternative can be calculated with the web calculator at *http://fpr-calc.ucl.ac.uk/*. However other approaches to the calculation of FPR, based on different

assumptions, give results that are similar (Table 1 in Colquhoun 2019).

To calculate FPR it is necessary to specify a prior probability and this is rarely known. My recommendation 2 is based on giving the FPR for a prior probability of 0.5. Any higher prior probability of there being a real effect is not justifiable in the absence of hard data. In this sense, the calculated FPR is the minimum that can be expected. More implausible hypotheses would make the problem worse. For example, if the prior probability of there being a real effect were only 0.1, then observation of $p = 0.05$ would imply a disastrously high FPR = 0.76, and in order to achieve an FPR of 0.05, you'd need to observe $p = 0.00045$. Others (especially Goodman) have advocated giving likelihood ratios (LRs) in place of p-values. The FPR for a prior of 0.5 is simply $1/(1 + \text{LR})$, so to give the FPR for a prior of 0.5 is simply a more-easily-comprehensible way of specifying the LR, and so should be acceptable to frequentists and Bayesians.

Matthews, R., Moving Toward the Post $p < 0.05$ Era via the Analysis of Credibility

1. Report the outcome of studies as effect sizes summarized by confidence intervals (CIs) along with their point estimates.
2. Make full use of the point estimate and width and location of the CI relative to the null effect line when interpreting findings. The point estimate is generally the effect size best supported by the study, irrespective of its statistical significance/nonsignificance. Similarly, tight CIs located far from the null effect line generally represent more compelling evidence for a nonzero effect than wide CIs lying close to that line.
3. Use the analysis of credibility (AnCred) to assess quantitatively the credibility of inferences based on the CI. AnCred determines the level of prior evidence needed for a new finding to provide credible evidence for a nonzero effect.
4. Establish whether this required level of prior evidence is supported by current knowledge and insight. If it is, the new result provides credible evidence for a nonzero effect, irrespective of its statistical significance/nonsignificance.

Gannon, M., Pereira, C., and Polpo, A., Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels

1. Retain the useful concept of statistical significance and the same operational procedures as currently used for hypothesis tests, whether frequentist (Neyman–Pearson p-value tests) or Bayesian (Bayes-factor tests).
2. Use tests with a sample-size-dependent significance level—ours is optimal in the sense of the generalized Neyman–Pearson lemma.
3. Use a testing scheme that allows tests of any kind of hypothesis, without restrictions on the dimensionalities of the parameter space or the hypothesis. Note that this should include "sharp" hypotheses, which correspond to subsets of lower dimensionality than the full parameter space.
4. Use hypothesis tests that are compatible with the likelihood principle (LP). They can be easier to interpret consistently than tests that are not LP-compliant.
5. Use numerical methods to handle hypothesis-testing problems with high-dimensional sample spaces or parameter spaces.

Pogrow, S., How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings

1. Switch from reliance on statistical or practical significance to the more stringent statistical criterion of practical benefit for (a) assessing whether applied research findings indicate that an intervention is effective and should be adopted and scaled—particularly in complex organizations such as schools and hospitals and (b) determining whether relationships are sufficiently strong and explanatory to be used as a basis for setting policy or practice recommendations. Practical benefit increases the likelihood that observed benefits will replicate in subsequent research and in clinical practice by avoiding the problems associated with relying on small effect sizes.
2. Reform statistics courses in applied disciplines to include the principles of practical benefit, and have students review influential applied research articles in the discipline to determine which findings demonstrate practical benefit.
3. Recognize the need to develop different inferential statistical criteria for assessing the importance of applied research findings as compared to assessing basic research findings.
4. Consider consistent, noticeable improvements across contexts using the quick prototyping methods of improvement science as a preferable methodology for identifying effective practices rather than on relying on RCT methods.
5. Require that applied research reveal the actual unadjusted means/medians of results for all groups and subgroups, and that review panels take such data into account—as opposed to only reporting relative differences between adjusted means/medians. This will help preliminarily identify whether there appear to be clear benefits for an intervention.

7.4. Adopting More Holistic Approaches

McShane, B., Gal, D., Gelman, A., Robert, C., and Tackett, J., Abandon Statistical Significance

1. Treat p-values (and other purely statistical measures like confidence intervals and Bayes factors) continuously rather than in a dichotomous or thresholded manner. In doing so, bear in mind that it seldom makes sense to calibrate evidence as a function of p-values or other purely statistical measures because they are, among other things, typically defined relative to the generally uninteresting and implausible null hypothesis of zero effect and zero systematic error.
2. Give consideration to related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain. Do this always—not just once some p-value or other statistical threshold has been attained—and do this without giving priority to p-values or other purely statistical measures.

3. Analyze and report all of the data and relevant results rather than focusing on single comparisons that attain some p-value or other statistical threshold.
4. Conduct a decision analysis: p-value and other statistical threshold-based rules implicitly express a particular tradeoff between Type I and Type II error, but in reality this tradeoff should depend on the costs, benefits, and probabilities of all outcomes.
5. Accept uncertainty and embrace variation in effects: we can learn much (indeed, more) about the world by forsaking the false promise of certainty offered by dichotomous declarations of truth or falsity—binary statements about there being "an effect" or "no effect"—based on some p-value or other statistical threshold being attained.
6. Obtain more precise individual-level measurements, use within-person or longitudinal designs more often, and give increased consideration to models that use informative priors, that feature varying treatment effects, and that are multilevel or meta-analytic in nature.

Tong, C., Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science

1. Prioritize effort for sound data production: the planning, design, and execution of the study.
2. Build scientific arguments with many sets of data and multiple lines of evidence.
3. Recognize the difference between exploratory and confirmatory objectives and use distinct statistical strategies for each.
4. Use flexible descriptive methodology, including disciplined data exploration, enlightened data display, and regularized, robust, and nonparametric models, for exploratory research.
5. Restrict statistical inferences to confirmatory analyses for which the study design and statistical analysis plan are prespecified prior to, and strictly adhered to during, data acquisition.

Amrhein, V., Trafimow, D., and Greenland, S., Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don't Expect Replication

1. Do not dichotomize, but embrace variation.
   (a) Report and interpret inferential statistics like the p-value in a continuous fashion; do not use the word "significant."
   (b) Interpret interval estimates as "compatibility intervals," showing effect sizes most compatible with the data, under the model used to compute the interval; do not focus on whether such intervals include or exclude zero.
   (c) Treat inferential statistics as highly unstable local descriptions of relations between models and the obtained data.
      (i) Free your "negative results" by allowing them to be potentially positive. Most studies with large p-values or interval estimates that include the null should be considered "positive," in the sense that they usually leave open the possibility of important effects (e.g., the effect sizes within the interval estimates).
      (ii) Free your "positive results" by allowing them to be different. Most studies with small p-values or interval estimates that are not near the null should be considered provisional, because in replication studies the p-values could be large and the interval estimates could show very different effect sizes.
      (iii) There is no replication crisis if we don't expect replication. Honestly reported results *must* vary from replication to replication because of varying assumption violations and random variation; excessive agreement itself would suggest deeper problems such as failure to publish results in conflict with group expectations.

Calin-Jageman, R., and Cumming, G., The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known

1. Ask quantitative questions and give quantitative answers.
2. Countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize, and interpret the potential for error.
3. Seek replication, and use quantitative methods to synthesize across data sets as a matter of course.
4. Use Open Science practices to enhance the trustworthiness of research results.
5. Avoid, wherever possible, any use of p-values or NHST.

Ziliak, S., How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little "p" Is Not Enough

- *G-10 Consider the Purpose of the Inquiry, and Compare with Best Practice.* Falsification of a null hypothesis is not the main purpose of the experiment or observational study. Making money or beer or medicine—ideally more and better than the competition and best practice—is. Estimating the importance of your coefficient relative to results reported by others, is. To repeat, as the 2016 ASA Statement makes clear, merely falsifying a null hypothesis with a qualitative yes/no, exists/does not exist, significant/not significant answer, is not itself significant science, and should be eschewed.
- *G-9 Estimate the Stakes (Or Eat Them).* Estimation of magnitudes of effects, and demonstrations of their substantive meaning, should be the center of most inquiries. Failure to specify the stakes of a hypothesis is the first step toward eating them (gulp).
- *G-8 Study Correlated Data: ABBA, Take a Chance on Me.* Most regression models assume "iid" error terms—independently and identically distributed—yet most data in the social and life sciences are correlated by systematic, nonrandom effects—and are thus not independent. Gosset solved the problem of correlated soil plots with the "ABBA" layout, maximizing the correlation of paired differences between the As and Bs with a perfectly balanced chiasmic arrangement.
- *G-7 Minimize "Real Error" with the 3 R's: Represent, Replicate, Reproduce.* A test of significance on a single set of data is nearly valueless. Fisher's p, Student's t, and other tests should only be used when there is actual repetition of the experi-

ment. "One and done" is scientism, not scientific. Random error is not equal to real error, and is usually smaller and less important than the sum of nonrandom errors. Measurement error, confounding, specification error, and bias of the auspices are frequently larger in all the testing sciences, agronomy to medicine. Guinnessometrics minimizes real error by repeating trials on stratified and balanced yet independent experimental units, controlling as much as possible for local fixed effects.

- *G-6 Economize with "Less is More": Small Samples of Independent Experiments.* Small sample analysis and distribution theory has an economic origin and foundation: changing inputs to the beer on the large scale (for Guinness, enormous global scale) is risky, with more than money at stake. But smaller samples, as Gosset showed in decades of barley and hops experimentation, does not mean "less than," and Big Data is in any case not the solution for many problems.
- *G-5 Keep Your Eyes on the Size Matters/How Much? Question.* There will be distractions but the expected loss and profit functions rule, or should. Are regression coefficients or differences between means large or small? Compared to what? How do you know?
- *G-4 Visualize.* Parameter uncertainty is not the same thing as model uncertainty. Does the result hit you between the eyes? Does the study show magnitudes of effects across the entire distribution? Advances in visualization software continue to outstrip advances in statistical modeling, making more visualization a no brainer.
- *G-3 Consider Posteriors and Priors too ("It pays to go Bayes").* The sample on hand is rarely the only thing that is "known." Subject matter expertise is an important prior input to statistical design and affects analysis of "posterior" results. For example, Gosset at Guinness was wise to keep quality assurance metrics and bottom line profit at the center of his inquiry. How does prior information fit into the story and evidence? Advances in Bayesian computing software make it easier and easier to do a Bayesian analysis, merging prior and posterior information, values, and knowledge.
- *G-2 Cooperate Up, Down, and Across (Networks and Value Chains).* For example, where would brewers be today without the continued cooperation of farmers? Perhaps back on the farm and not at the brewery making beer. Statistical science is social, and cooperation helps. Guinness financed a large share of modern statistical theory, and not only by supporting Gosset and other brewers with academic sabbaticals (Ziliak and McCloskey 2008).
- *G-1 Answer the Brewer's Original Question ("How should you set the odds?").* No bright-line rule of statistical significance can answer the brewer's question. As Gosset said way back in 1904, how you set the odds depends on "the importance of the issues at stake" (e.g., the expected benefit and cost) together with the cost of obtaining new material.

Billheimer, D., Predictive Inference and Scientific Reproducibility

1. Predict observable events or quantities that you care about.
2. Quantify the uncertainty of your predictions.

Manski, C., Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing

1. Statisticians should relearn statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten by the century's end.
2. Statistical decision theory should supplant hypothesis testing when statisticians study treatment choice with trial data.
3. Statisticians should use statistical decision theory when analyzing decision making with sample data more generally.

Manski, C., and Tetenov, A., Trial Size for Near Optimal Choice between Surveillance and Aggressive Treatment: Reconsidering MSLT-II

1. Statisticians should relearn statistical decision theory, which received considerable attention in the middle of the twentieth century but was largely forgotten by the century's end.
2. Statistical decision theory should supplant hypothesis testing when statisticians study treatment choice with trial data.
3. Statisticians should use statistical decision theory when analyzing decision making with sample data more generally.

Lavine, M., Frequentist, Bayes, or Other?

1. Look for and present results from many models that fit the data well.
2. Evaluate models, not just procedures.

Ruberg, S., Harrell, F., Gamalo-Siebers, M., LaVange, L., Lee J., Price K., and Peck C., Inference and Decision-Making for 21st Century Drug Development and Approval

1. Apply Bayesian paradigm as a framework for improving statistical inference and regulatory decision making by using probability assertions about the magnitude of a treatment effect.
2. Incorporate prior data and available information formally into the analysis of the confirmatory trials.
3. Justify and pre-specify how priors are derived and perform sensitivity analysis for a better understanding of the impact of the choice of prior distribution.
4. Employ quantitative utility functions to reflect key considerations from all stakeholders for optimal decisions via a probability-based evaluation of the treatment effects.
5. Intensify training in Bayesian approaches, particularly for decision makers and clinical trialists (e.g., physician scientists in FDA, industry and academia).

van Dongen, N., Wagenmakers, E.J., van Doorn, J., Gronau, Q., van Ravenzwaaij, D., Hoekstra, R., Haucke, M., Lakens, D., Hennig, C., Morey, R., Homer, S., Gelman, A., and Sprenger, J., Multiple Perspectives on Inference for Two Simple Statistical Scenarios

1. Clarify your statistical goals explicitly and unambiguously.
2. Consider the question of interest and choose a statistical approach accordingly.
3. Acknowledge the uncertainty in your statistical conclusions.
4. Explore the robustness of your conclusions by executing several different analyses.
5. Provide enough background information such that other researchers can interpret your results and possibly execute meaningful alternative analyses.

7.5. Reforming Institutions: Changing Publication Policies and Statistical Education

Trafimow, D., Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post $P < 0.05$ Universe

1. Tolerate ambiguity.
2. Replace significance testing with a priori thinking.
3. Consider the nature of the contribution, on multiple levels.
4. Emphasize thinking and execution, not results.
5. Consider that the assumption of random and independent sampling might be wrong.

Locascio, J., The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration

For journal reviewers

1. Provide an initial provisional decision regarding acceptance for publication of a journal manuscript based exclusively on the judged importance of the research issues addressed by the study and the soundness of the reported methodology. (The latter would include appropriateness of data analysis methods.) Give no weight to the reported results of the study per se in the decision as to whether to publish or not.
2. To ensure #1 above is accomplished, commit to an initial decision regarding publication after having been provided with only the Introduction and Methods sections of a manuscript by the editor, not having seen the Abstract, Results, or Discussion. (The latter would be reviewed only if and after a generally irrevocable decision to publish has already been made.)

For investigators/manuscript authors

1. Obtain consultation and collaboration from statistical consultant(s) and research methodologist(s) early in the development and conduct of a research study.
2. Emphasize the clinical and scientific importance of a study in the Introduction section of a manuscript, and give a clear, explicit statement of the research questions being addressed and any hypotheses to be tested.
3. Include a detailed statistical analysis subsection in the Methods section, which would contain, among other things, a justification of the adequacy of the sample size and the reasons various statistical methods were employed. For example, if null hypothesis significance testing and p-values are used, presumably supplemental to other methods, justify why those methods apply and will provide useful additional information in this particular study.
4. Submit for publication reports of well-conducted studies on important research issues regardless of findings, for example, even if only null effects were obtained, hypotheses were not confirmed, mere replication of previous results were found, or results were inconsistent with established theories.

Hurlbert, S., Levine, R., and Utts, J., Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires

1. Encourage journal editorial boards to disallow use of the phrase "statistically significant," or even "significant," in manuscripts they will accept for review.
2. Give primary emphasis in abstracts to the magnitudes of those effects most conclusively demonstrated and of greatest import to the subject matter.
3. Report precise p-values or other indices of evidence against null hypotheses as continuous variables not requiring any labeling.
4. Understand the meaning of and rationale for neoFisherian significance assessment (NFSA).

Campbell, H., and Gustafson, P., The World of Research Has Gone Berserk: Modeling the Consequences of Requiring "Greater Statistical Stringency" for Scientific Publication

1. *Consider the meta-research implications of implementing new publication/funding policies.* Journal editors and research funders should attempt to model the impact of proposed policy changes before any implementation. In this way, we can anticipate the policy impacts (both positive and negative) on the types of studies researchers pursue and the types of scientific articles that ultimately end up published in the literature.

Fricker, R., Burke, K., Han, X., and Woodall, W., Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their p-Value Ban

1. Use measures of statistical significance combined with measures of practical significance, such as confidence intervals on effect sizes, in assessing research results.
2. Classify research results as either exploratory or confirmatory and appropriately describe them as such in all published documentation.
3. Define precisely the population of interest in research studies and carefully assess whether the data being analyzed are representative of the population.
4. Understand the limitations of inferential methods applied to observational, convenience, or other nonprobabilistically sampled data.

Maurer, K., Hudiburgh, L., Werwinski, L., and Bailer J., Content Audit for p-Value Principles in Introductory Statistics

1. Evaluate the coverage of p-value principles in the introductory statistics course using rubrics or other systematic assessment guidelines.
2. Discuss and deploy improvements to curriculum coverage of p-value principles.
3. Meet with representatives from other departments, who have majors taking your statistics courses, to make sure that inference is being taught in a way that fits the needs of their disciplines.
4. Ensure that the correct interpretation of p-value principles is a point of emphasis for all faculty members and embedded within all courses of instruction.

Steel, A., Liermann, M., and Guttorp, P., Beyond Calculations: A Course in Statistical Thinking

1. Design curricula to teach students how statistical analyses are embedded within a larger science life-cycle, including steps such as project formulation, exploratory graphing, peer review, and communication beyond scientists.
2. Teach the p-value as only one aspect of a complete data analysis.

3. Prioritize helping students build a strong understanding of what testing and estimation can tell you over teaching statistical procedures.
4. Explicitly teach statistical communication. Effective communication requires that students clearly formulate the benefits and limitations of statistical results.
5. Force students to struggle with poorly defined questions and real, messy data in statistics classes.
    5. Encourage students to match the mathematical metric (or data summary) to the scientific question. Teaching students to create customized statistical tests for custom metrics allows statistics to move beyond the mean and pinpoint specific scientific questions.

Acknowledgments

Without the help of a huge team, this special issue would never have happened. The articles herein are about the equivalent of three regular issues of *The American Statistician*. Thank you to all the authors who submitted papers for this issue. Thank you, authors whose papers were accepted, for enduring our critiques. We hope they made you happier with your finished product. Thank you to a talented, hard-working group of associate editors for handling many papers: Frank Bretz, George Cobb, Doug Hubbard, Ray Hubbard, Michael Lavine, Fan Li, Xihong Lin, Tom Louis, Regina Nuzzo, Jane Pendergast, Annie Qu, Sherri Rose, and Steve Ziliak. Thank you to all who served as reviewers. We definitely couldn't have done this without you. Thank you, TAS Editor Dan Jeske, for your vision and your willingness to let us create this special issue. Special thanks to Janet Wallace, TAS editorial coordinator, for spectacular work and tons of patience. We also are grateful to ASA Journals Manager Eric Sampson for his leadership, and to our partners, the team at Taylor and Francis, for their commitment to ASA's publishing efforts. Thank you to all who read and commented on the draft of this editorial. You made it so much better! Regina Nuzzo provided extraordinarily helpful substantive and editorial comments. And thanks most especially to the ASA Board of Directors, for generously and enthusiastically supporting the "*p*-values project" since its inception in 2014. Thank you for your leadership of our profession and our association.

Gratefully,
Ronald L. Wasserstein
American Statistical Association, Alexandria, VA
ron@amstat.org

Allen L. Schirm
Mathematica Policy Research (retired), Washington, DC
allenschirm@gmail.com

Nicole A. Lazar
Department of Statistics, University of Georgia, Athens, GA
nlazar@stat.uga.edu

References

References to articles in this special issue

Amrhein, V., Trafimow, D., and Greenland, S. (2019), "Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don't Expect Replication," *The American Statistician*, 73. [2,3,4,5,6,7,8,9]

Anderson, A. (2019), "Assessing Statistical Results: Magnitude, Precision and Model Uncertainty," *The American Statistician*, 73. [3]

Benjamin, D., and Berger, J. (2019), "Three Recommendations for Improving the Use of *p*-Values," *The American Statistician*, 73. [5,7,9]

Betensky, R. (2019), "The *p*-Value Requires Context, Not a Threshold," *The American Statistician*, 73. [4,6]

Billheimer, D. (2019), "Predictive Inference and Scientific Reproducibility," *The American Statistician*, 73. [5]

Blume, J., Greevy, R., Welty, V., Smith, J., and DuPont, W. (2019), "An Introduction to Second Generation *p*-Value," *The American Statistician*, 73. [4]

Brownstein, N., Louis, T., O'Hagan, A., and Pendergast, J. (2019), "The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making," *The American Statistician*, 73. [5]

Calin-Jageman, R., and Cumming, G. (2019), "The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known," *The American Statistician*, 73. [3,5,9,10]

Campbell, H., and Gustafson, P. (2019), "The World of Research Has Gone Berserk: Modeling the Consequences of Requiring 'Greater Statistical Stringency' for Scientific Publication," *The American Statistician*, 73. [8]

Colquhoun, D. (2019), "The False Positive Risk: A Proposal Concerning What to Do About *p*-Value," *The American Statistician*, 73. [4,7,14]

Fraser, D. (2019), "The *p*-Value Function and Statistical Inference," *The American Statistician*, 73. [6]

Fricker, R., Burke, K., Han, X., and Woodall, W (2019), "Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their *p*-Value Ban," *The American Statistician*, 73. [7,9]

Gannon, M., Pereira, C., and Polpo, A. (2019), "Blending Bayesian and Classical Tools to Define Optimal Sample-Size-Dependent Significance Levels," *The American Statistician*, 73. [5]

Goodman, S. (2019), "Why is Getting Rid of *p*-Values So Hard? Musings on Science and Statistics," *The American Statistician*, 73. [7,8,10]

Goodman, W., Spruill, S., and Komaroff, E. (2019), "A Proposed Hybrid Effect Size Plus *p*-Value Criterion: Empirical Evidence Supporting Its Use," *The American Statistician*, 73. [5]

Greenland, S. (2019), "Valid *p*-Values Behave Exactly as They Should: Some Misleading Criticisms of *p*-Values and Their Resolution With *s*-Values," *The American Statistician*, 73. [3,4,5,6,7]

Heck, P., and Krueger, J. (2019), "Putting the *p*-Value in Its Place," *The American Statistician*, 73. [4]

Hubbard, D., and Carriquiry, A. (2019), "Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness and Relevance," *The American Statistician*, 73. [5]

Hubbard, R. (2019), "Will the ASA's Efforts to Improve Statistical Practice Be Successful? Some Evidence to the Contrary," *The American Statistician*, 73. [8]

Hubbard, R., Haig, B. D., and Parsa, R. A. (2019), "The Limited Role of Formal Statistical Inference in Scientific Inference," *The American Statistician*, 73. [2,7]

Hurlbert, S., Levine, R., and Utts, J. (2019), "Coup de Grâce for a Tough Old Bull: 'Statistically Significant' Expires," *The American Statistician*, 73. [8,9]

Ioannidis, J. (2019), "What Have We (Not) Learnt From Millions of Scientific Papers With *p*-Values?," *The American Statistician*, 73. [6]

Johnson, V. (2019), "Evidence From Marginally Significant t Statistics," *The American Statistician*, 73. [7]

Kennedy-Shaffer, L. (2019), "Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize *p*-Values and Significance Testing," *The American Statistician*, 73. [7]

Kmetz, J. (2019), "Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of *p*-Values," *The American Statistician*, 73. [8]

Lavine, M. (2019), "Frequentist, Bayes, or Other?," *The American Statistician*, 73. [6]

Locascio, J. (2019), "The Impact of Results Blind Science Publishing on Statistical Consultation and Collaboration," *The American Statistician*, 73. [4,5,8,9]

Manski, C. (2019), "Treatment Choice With Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing," *The American Statistician*, 73. [5]

Manski, C., and Tetenov, A. (2019), "Trial Size for Near Optimal Choice between Surveillance and Aggressive Treatment: Reconsidering MSLT-II," *The American Statistician*, 73. [5]

Matthews, R. (2019), "Moving Toward the Post $p < 0.05$ Era Via the Analysis of Credibility," *The American Statistician*, 73. [4,9]

Maurer, K., Hudiburgh, L., Werwinski, L., and Bailer, J. (2019), "Content Audit for *P*-Value Principles in Introductory Statistics," *The American Statistician*, 73. [8]

McShane, B., Gal, D., Gelman, A., Robert, C., and Tackett, J. (2019), "Abandon Statistical Significance," *The American Statistician*, 73. [4,6,7]

McShane, B., Tackett, J., Böckenholt, U., and Gelman, A. (2019), "Large Scale Replication Projects in Contemporary Psychological Research," *The American Statistician*, 73. [9]

O'Hagan, A. (2019), "Expert Knowledge Elicitation: Subjective But Scientific," *The American Statistician*, 73. [5,6]

Pogrow, S. (2019), "How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings," *The American Statistician*, 73. [4]

Rose, S., and McGuire, T. (2019), "Limitations of *p*-Values and *R*-Squared for Stepwise Regression Building: A Fairness Demonstration in Health Policy Risk Adjustment," *The American Statistician*, 73. [7]

Rougier, J. (2019), "*p*-Values, Bayes Factors, and Sufficiency," *The American Statistician*, 73. [6]

Ruberg, S., Harrell, F., Gamalo-Siebers, M., LaVange, L., Lee J., Price K., and Peck C. (2019), "Inference and Decision-Making for 21st Century Drug Development and Approval," *The American Statistician*, 73. [10]

Steel, A., Liermann, M., and Guttorp, P. (2019), "Beyond Calculations: A Course in Statistical Thinking," *The American Statistician*, 73. [8]

Trafimow, D. (2019), "Five Nonobvious Changes in Editorial Practice for Editors and Reviewers to Consider When Evaluating Submissions in a Post $p < .05$ Universe," *The American Statistician*, 73. [7]

Tong, C. (2019), "Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science," *The American Statistician*, 73. [2,3,4,6,8,9]

van Dongen, N., Wagenmakers, E. J., van Doorn, J., Gronau, Q., van Ravenzwaaij, D., Hoekstra, R., Haucke, M., Lakens, D., Hennig, C., Morey, R., Homer, S., Gelman, A., and Sprenger, J. (2019), "Multiple Perspectives on Inference for Two Simple Statistical Scenarios," *The American Statistician*, 73. [6]

Ziliak, S. (2019), "How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little 'P' is Not Enough," *The American Statistician*, 73. [2,3]

Other articles or books referenced

Boring, E. G. (1919), "Mathematical vs. Scientific Significance," *Psychological Bulletin*, 16, 335–338. [2]

Cumming, G. (2014), "The New Statistics: Why and How," *Psychological Science*, 25, 7–29. [3]

Davidian, M., and Louis, T. (2012), "Why Statistics?" *Science*, 336, 12. [2]

Edgeworth, F. Y. (1885), "Methods of Statistics," *Journal of the Statistical Society of London*, Jubilee Volume, 181–217. [2]

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd. [2]

Gelman, A. (2015), "Statistics and Research Integrity," *European Science Editing*, 41, 13–14. [8]

——— (2016), "The Problems With *p*-Values Are Not Just With *p*-Values," *The American Statistician*, supplemental materials to *ASA Statement on p-Values and Statistical Significance*, 70, 1–2. [3]

Gelman, A., and Hennig, C. (2017), "Beyond Subjective and Objective in Statistics," *Journal of the Royal Statistical Society*, Series A, 180, 967–1033. [5]

Gelman, A., and Stern, H. (2006), "The Difference Between 'Significant' and 'Not Significant' Is Not Itself Statistically Significant," *The American Statistician*, 60, 328–331. [2]

Ghose, T. (2013), "'Just a Theory': 7 Misused Science Words," *Scientific American* (online), available at *https://www.scientificamerican.com/article/just-a-theory-7-misused-science-words/*. [2]

Goodman, S. (2018), "How Sure Are You of Your Result? Put a Number on It," *Nature*, 564. [5]

Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Thousand Oaks, CA: Sage. [1]

Junger, S. (1997), *The Perfect Storm: A True Story of Men Against the Sea*, New York: W.W. Norton. [10]

Locascio, J. (2017), "Results Blind Science Publishing," *Basic and Applied Social Psychology*, 39, 239–246. [8]

Mayo, D. (2018), "Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars," Cambridge, UK: University Printing House. [1]

McShane, B., and Gal, D. (2016), "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence," *Management Science*, 62, 1707–1718. [9]

——— (2017), "Statistical Significance and the Dichotomization of Evidence." *Journal of the American Statistical Association*, 112, 885–895. [9]

Mogil, J. S., and Macleod, M. R. (2017), "No Publication Without Confirmation," *Nature*, 542, 409–411, available at *https://www.nature.com/news/no-publication-without-confirmation-1.21509*. [8]

Rosenthal, R. (1979), "File Drawer Problem and Tolerance for Null Results," *Psychological Bulletin* 86, 638–641. [2]

Wasserstein, R., and Lazar, N. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [1]

Wellek, S. (2017), "A Critical Evaluation of the Current *p*-Value Controversy" (with discussion), *Biometrical Journal*, 59, 854–900. [4,9]

Ziliak, S., and McCloskey, D. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [1,16]

Exhibit 63

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://amstat.tandfonline.com/loi/utas20

The ASA's Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on *p*-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

To link to this article: https://doi.org/10.1080/00031305.2016.1154108

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews (Siegfried 2010) wrote: "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a fiimsy foundation." A November 2013, article in Phys.org Science News Wire (2013) cited "numerous deep fiaws" in null hypothesis significance testing. A ScienceNews article (Siegfried 2014) on February 7, 2014, said "statistical techniques for testing hypotheses …have more fiaws than Facebook's privacy policies." A week later, statistician and "Simply Statistics" blogger Jefi Leek responded. "The problem is not that people use P-values poorly," Leek wrote, "it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis" (Leek 2014). That same week, statistician and science writer Regina Nuzzo published an article in *Nature* entitled "Scientific Method: Statistical Errors" (Nuzzo 2014). That article is now one of the most highly viewed *Nature* articles, as reported by altmetric.com (*http://www.altmetric.com/details/2115792#score*).

Of course, it was not simply a matter of responding to some articles in print. The statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions. Without getting into definitions and distinctions of these terms, we observe that much confusion and even doubt about the validity of science is arising. Such doubt can lead to radical choices, such as the one taken by the editors of *Basic and Applied Social Psychology*, who decided to ban *p*-values (null hypothesis significance testing) (Trafimow and Marks 2015). Misunderstanding or misuse of statistical inference is only one cause of the "reproducibility crisis" (Peng 2015), but to our community, it is an important one.

When the ASA Board decided to take up the challenge of developing a policy statement on *p*-values and statistical significance, it did so recognizing this was not a lightly taken step. The ASA has not previously taken positions on specific matters of statistical practice. The closest the association has come to this is a statement on the use of value-added models (VAM) for educational assessment (Morganstein and Wasserstein 2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as efiective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on *p*-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provides the community a service. The intended audience would be researchers, practitioners, and science writers who are not primarily statisticians. Thus, this statement would be quite difierent from anything previously attempted.

The Board tasked Wasserstein with assembling a group of experts representing a wide variety of points of view. On behalf of the Board, he reached out to more than two dozen such people, all of whom said they would be happy to be involved. Several expressed doubt about whether agreement could be reached, but those who did said, in efiect, that if there was going to be a discussion, they wanted to be involved.

Over the course of many months, group members discussed what format the statement should take, tried to more concretely visualize the audience for the statement, and began to find points of agreement. That turned out to be relatively easy to do, but it was just as easy to find points of intense disagreement.

The time came for the group to sit down together to hash out these points, and so in October 2015, 20 members of the group met at the ASA Ofi ce in Alexandria, Virginia. The 2-day meeting was facilitated by Regina Nuzzo, and by the end of the meeting, a good set of points around which the statement could be built was developed.

The next 3 months saw multiple drafts of the statement, reviewed by group members, by Board members (in a lengthy discussion at the November 2015 ASA Board meeting), and by members of the target audience. Finally, on January 29, 2016, the Executive Committee of the ASA approved the statement.

The statement development process was lengthier and more controversial than anticipated. For example, there was considerable discussion about how best to address the issue of multiple *potential* comparisons (Gelman and Loken 2014). We debated at some length the issues behind the words "a *p*-value near 0.05 taken by itself ofiers only weak evidence against the null

hypothesis" (Johnson 2013). There were difiering perspectives about how to characterize various alternatives to the *p*-value and in how much detail to address them. To keep the statement reasonably simple, we did not address alternative hypotheses, error types, or power (among other things), and not everyone agreed with that approach.

As the end of the statement development process neared, Wasserstein contacted Lazar and asked if the policy statement might be appropriate for publication in *The American Statistician* (*TAS*). After consideration, Lazar decided that *TAS* would provide a good platform to reach a broad and general statistical readership. Together, we decided that the addition of an online discussion would heighten the interest level for the *TAS* audience, giving an opportunity to refiect the aforementioned controversy.

To that end, a group of discussants was contacted to provide comments on the statement. You can read their statements in the online supplement, and a guide to those statements appears at the end of this editorial. We thank Naomi Altman, Douglas Altman, Daniel J. Benjamin, Yoav Benjamini, Jim Berger, Don Berry, John Carlin, George Cobb, Andrew Gelman, Steve Goodman, Sander Greenland, John Ioannidis, Joseph Horowitz, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Michele Millar, Charles Poole, Ken Rothman, Stephen Senn, Dalene Stangl, Philip Stark and Steve Ziliak for sharing their insightful perspectives.

Of special note is the following article, which is a significant contribution to the literature about *p*-values and statistical significance.

> Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. and Altman, D.G.: "Statistical Tests, *P*-values, Confidence Intervals, and Power: A Guide to Misinterpretations."

Though there was disagreement on exactly what the statement should say, there was high agreement that the ASA should be speaking out about these matters.

Let us be clear. Nothing in the ASA statement is new. Statisticians and others have been sounding the alarm about these matters for decades, to little avail. We hoped that a statement from the world's largest professional association of statisticians would open a fresh discussion and draw renewed and vigorous attention to changing the practice of science with regards to the use of statistical inference.

Guide to the Online Supplemental Material to the ASA Statement on *P*-Values and Statistical Significance

Many of the participants in the development of the ASA statement contributed commentary about the statement or matters related to it. Their comments are posted as online supplements to the statement. We provide here a list of the supplemental articles.

Supplemental Material to the ASA Statement on *P*-Values and Statistical Significance

- *Altman, Naomi*: Ideas from multiple testing of high dimensional data provide insights about reproducibility and false discovery rates of hypothesis supported by *p*-values
- *Benjamin, Daniel J, and Berger, James O:* A simple alternative to *p*-values
- *Benjamini, Yoav:* It's not the *p*-values' fault
- *Berry, Donald A: P*-values are not what they're cracked up to be
- *Carlin, John B:* Comment: Is reform possible without a paradigm shift?
- *Cobb, George:* ASA statement on p-values: Two consequences we can hope for
- *Gelman, Andrew:* The problems with *p*-values are not just with *p*-values
- *Goodman, Steven N:* The next questions: Who, what, when, where, and why?
- *Greenland, Sander:* The ASA guidelines and null bias in current teaching and practice
- *Ioannidis, John P.A.:* Fit-for-purpose inferential methods: abandoning/changing *P*-values versus abandoning/changing research
- *Johnson, Valen E.:* Comments on the "ASA Statement on Statistical Significance and *P*-values" and marginally significant *p*-values
- *Lavine, Michael, and Horowitz, Joseph:* Comment
- *Lew, Michael J:* Three inferential questions, two types of *P*-value
- *Little, Roderick J:* Discussion
- *Mayo, Deborah G:* Don't throw out the error control baby with the bad statistics bathwater
- *Millar, Michele:* ASA statement on *p*-values: some implications for education
- *Rothman, Kenneth J:* Disengaging from statistical significance
- *Senn, Stephen:* Are *P*-Values the Problem?
- *Stangl, Dalene:* Comment
- *Stark, P.B.:* The value of *p*-values
- *Ziliak, Stephen T*: The significance of the ASA statement on statistical significance and *p*-values

References

American Statistical Association (2010), "ASA Statement on Risk-Limiting Post Election Audits." Available at *http://www.amstat.org/policy/pdfs/Risk-Limiting_Endorsement.pdf*. [129]

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science [online]," *American Scientist*, 102. Available at *http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science*. [129]

Johnson, V. E. (2013), "Uniformly Most Powerful Bayesian Tests," *Annals of Statistics*, 41, 1716–1741. [130]

Leek, J. (2014), "On the Scalability of Statistical Procedures: Why the *p*-Value Bashers Just Don't Get It," Simply Statistics Blog, Available at *http://simplystatistics.org/2014/02/14/on-the-scalability-of-statistical-procedures-why-the-p-value-bashers-just-dont-get-it/*. [129]

Morganstein, D., and Wasserstein, R. (2014), "ASA Statement on Value-Added Models," *Statistics and Public Policy*, 1, 108–110. Available at *http://amstat.tandfonline.com/doi/full/10.1080/2330443X.2014.956906*. [129]

Nuzzo, R. (2014), "Scientific Method: Statistical Errors," *Nature*, 506, 150–152. Available at *http://www.nature.com/news/scientific-method-statistical-errors-1.14700*. [129]

Peng, R. (2015), "The Reproducibility Crisis in Science: A Statistical Counterattack," *Significance*, 12, 30–32. [129]

Phys.org Science News Wire (2013), "The Problem With *p* Values: How Significant are They, Really?" Available at *http://phys.org/wire-news/*

145707973/the-problem-with-p-values-how-significant-are-they-really.html. [129]

Siegfried, T. (2010), "Odds Are, It's Wrong: Science Fails to Face the Shortcomings of Statistics," *Science News*, 177, 26. Available at *https://www.sciencenews.org/article/odds-are-its-wrong*. [129]

Siegfried, T. (2014), "To Make Science Better, Watch out for Statistical Flaws," *Science News Context Blog*, February 7, 2014. Available at *https://www.sciencenews.org/blog/context/make-science-better-watch-out-statistical-flaws*. [129]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology* 37, 1–2. [129]

Ronald L. Wasserstein and Nicole A. Lazar
ron@amstat.org
American Statistical Association, 732 North Washington Street, Alexandria, VA 22314-1943.

ASA Statement on Statistical Significance and *P*-Values

1. Introduction

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the *p*-value. While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of *p*-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since *p*-values were first introduced.

In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in nontechnical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

2. What is a *p*-Value?

Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

3. Principles

1. ***P*-values can indicate how incompatible the data are with a specified statistical model.**

   A *p*-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. ***P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**

   Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.**

   Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

4. **Proper inference requires full reporting and transparency**

   P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the

reported *p*-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "*p*-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all *p*-values computed. Valid scientific conclusions based on *p*-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including *p*-values) were selected for reporting.

5. **A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.**

   Statistical significance is not equivalent to scientific, human, or economic significance. Smaller *p*-values do not necessarily imply the presence of larger or more important effects, and larger *p*-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small *p*-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive *p*-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different *p*-values if the precision of the estimates differs.

6. **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.**

   Researchers should recognize that a *p*-value without context or other evidence provides limited information. For example, a *p*-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large *p*-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a *p*-value when other approaches are appropriate and feasible.

4. Other Approaches

In view of the prevalent misuses of and misconceptions concerning *p*-values, some statisticians prefer to supplement or even replace *p*-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct.

5. Conclusion

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

Acknowledgments

The ASA Board of Directors thanks the following people for sharing their expertise and perspectives during the development of the statement. The statement does not necessarily reflect the viewpoint of all these people, and in fact some have views that are in opposition to all or part of the statement. Nonetheless, we are deeply grateful for their contributions. Naomi Altman, Jim Berger, Yoav Benjamini, Don Berry, Brad Carlin, John Carlin, George Cobb, Marie Davidian, Steve Fienberg, Andrew Gelman, Steve Goodman, Sander Greenland, Guido Imbens, John Ioannidis, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Chuck McCulloch, Michele Millar, Sally Morton, Regina Nuzzo, Hilary Parker, Kenneth Rothman, Don Rubin, Stephen Senn, Uri Simonsohn, Dalene Stangl, Philip Stark, Steve Ziliak.

Edited by Ronald L. Wasserstein, Executive Director
On behalf of the American Statistical Association Board of Directors

A Brief *p*-Values and Statistical Significance Reference List

Altman D.G., and Bland J.M. (1995), "Absence of Evidence is not Evidence of Absence," *British Medical Journal*, 311, 485.

Altman, D.G., Machin, D., Bryant, T.N., and Gardner, M.J. (eds.) (2000), *Statistics with Confidence* (2nd ed.), London: BMJ Books.

Berger, J.O., and Delampady, M. (1987), "Testing Precise Hypotheses," *Statistical Science*, 2, 317–335.

Berry, D. (2012), "Multiplicities in Cancer Research: Ubiquitous and Necessary Evils," *Journal of the National Cancer Institute*, 104, 1124–1132.

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," *The American Statistician*, 59, 121–126.

Cox, D.R. (1982), "Statistical Significance Tests," *British Journal of Clinical Pharmacology*, 14, 325–331.

Edwards, W., Lindman, H., and Savage, L.J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science [online]," *American Scientist*, 102. Available at *http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science*

Gelman, A., and Stern, H.S. (2006), "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant," *The American Statistician*, 60, 328–331.

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socioeconomics*, 33, 567–606.

Goodman, S.N. (1999a), "Toward Evidence-Based Medical Statistics 1: The *P*-Value Fallacy," *Annals of Internal Medicine*, 130, 995–1004.

——— (1999b), "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor," *Annals of Internal Medicine*, 130, 1005–1013.

——— (2008), "A Dirty Dozen: Twelve *P*-Value Misconceptions," *Seminars in Hematology*, 45, 135–140.

Greenland, S. (2011), "Null Misinterpretation in Statistical Testing and its Impact on Health Risk Assessment," *Preventive Medicine*, 53, 225–228.

——— (2012), "Nonsignificance Plus High Power Does Not Imply Support for the Null Over the Alternative," *Annals of Epidemiology*, 22, 364–368.

Greenland, S., and Poole, C. (2011), "Problems in Common Interpretations of Statistics in Scientific Articles, Expert Reports, and Testimony," *Jurimetrics*, 51, 113–129.

Hoenig, J.M., and Heisey, D.M. (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician*, 55, 19–24.

Ioannidis, J.P. (2005), "Contradicted and Initially Stronger Efiects in Highly Cited Clinical Research," *Journal of the American Medical Association*, 294, 218–228.

——— (2008), "Why Most Discovered True Associations are Inflated" (with discussion), *Epidemiology* 19, 640–658.

Johnson, V.E. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences*, 110(48), 19313–19317.

——— (2013), "Uniformly Most Powerful Bayesian Tests," *Annals of Statistics*, 41, 1716–1741.

Lang, J., Rothman K.J., and Cann, C.I. (1998), "That Confounded *P*-value" (editorial), *Epidemiology*, 9, 7–8.

Lavine, M. (1999), "What is Bayesian Statistics and Why Everything Else is Wrong," *UMAP Journal*, 20, 2.

Lew, M.J. (2012), "Bad Statistical Practice in Pharmacology (and Other Basic Biomedical Disciplines): You Probably Don't Know *P*," *British Journal of Pharmacology*, 166, 5, 1559–1567.

Phillips, C.V. (2004), "Publication Bias In Situ," *BMC Medical Research Methodology*, 4, 20.

Poole, C. (1987), "Beyond the Confidence Interval," *American Journal of Public Health*, 77, 195–199.

——— (2001), "Low *P*-values or Narrow Confidence Intervals: Which are More Durable?" *Epidemiology*, 12, 291–294.

Rothman, K.J. (1978), "A Show of Confidence" (editorial), *New England Journal of Medicine*, 299, 1362–1363.

——— (1986), "Significance Questing" (editorial), *Annals of Internal Medicine*, 105, 445–447.

———- (2010), "Curbing Type I and Type II Errors," *European Journal of Epidemiology*, 25, 223–224.

Rothman, K.J., Weiss, N.S., Robins, J., Neutra, R., and Stellman, S. (1992), "Amicus Curiae Brief for the U. S. Supreme Court, *Daubert v. Merrell Dow Pharmaceuticals*, Petition for Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit," No. 92-102, October Term, 1992.

Rozeboom, W.M. (1960), "The Fallacy of the Null-Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–428.

Schervish, M.J. (1996), "*P*-Values: What They Are and What They Are Not," *The American Statistician*, 50, 203–206.

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366.

Stang, A., and Rothman, K.J. (2011), "That Confounded *P*-value Revisited," *Journal of Clinical Epidemiology*, 64, 1047–1048.

Stang, A., Poole, C., and Kuss, O. (2010), "The Ongoing Tyranny of Statistical Significance Testing in Biomedical Research," *European Journal of Epidemiology*, 25, 225–230.

Sterne, J. A. C. (2002). "Teaching Hypothesis Tests—Time for Significant Change?" *Statistics in Medicine*, 21, 985–994.

Sterne, J. A. C., and Smith, G. D. (2001), "Sifting the Evidence—What's Wrong with Significance Tests?" *British Medical Journal*, 322, 226–231.

Ziliak, S.T. (2010), "The Validus Medicus and a New Gold Standard," *The Lancet*, 376, 9738, 324–325.

Ziliak, S.T., and McCloskey, D.N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press.

Exhibit 64

THIRD EDITION

MODERN EPIDEMIOLOGY

Kenneth J. Rothman
Vice President, Epidemiology Research
RTI Health Solutions
Professor of Epidemiology and Medicine
Boston University
Boston, Massachusetts

Sander Greenland
Professor of Epidemiology and Statistics
University of California
Los Angeles, California

Timothy L. Lash
Associate Professor of Epidemiology and Medicine
Boston University
Boston, Massachusetts

Wolters Kluwer Health | Lippincott Williams & Wilkins
Philadelphia • Baltimore • New York • London
Buenos Aires • Hong Kong • Sydney • Tokyo

Acquisitions Editor: Sonya Seigafuse
Developmental Editor: Louise Bierig
Project Manager: Kevin Johnson
Senior Manufacturing Manager: Ben Rivera
Marketing Manager: Kimberly Schonberger
Art Director: Risa Clow
Compositor: Aptara, Inc.

530 Walnut Street
Philadelphia, PA 19106 USA
LWW.com

Printed in the USA

Library of Congress Cataloging-in-Publication Data
Rothman, Kenneth J.
Modern epidemiology / Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. – 3rd ed.
p. ; cm.
2nd ed. edited by Kenneth J. Rothman and Sander Greenland.
Includes bibliographical references and index.
ISBN-13: 978-0-7817-5564-1
ISBN-10: 0-7817-5564-6
1. Epidemiology–Statistical methods. 2. Epidemiology–Research–Methodology. I. Greenland, Sander, 1951- II. Lash, Timothy L. III. Title.
[DNLM: 1. Epidemiology. 2. Epidemiologic Methods. WA 105 R846m 2008]
RA652.2.M3R67 2008
614.4–dc22 2007036316

10 9 8 7 6 5 4 3 2 1

CHAPTER 2

Causation and Causal Inference

Kenneth J. Rothman, Sander Greenland,
Charles Poole, and Timothy L. Lash

CAUSALITY

A rudimentary understanding of cause and effect seems to be acquired by most people on their own much earlier than it could have been taught to them by someone else. Even before they can speak, many youngsters understand the relation between crying and the appearance of a parent or other adult, and the relation between that appearance and getting held, or fed. A little later, they will develop theories about what happens when a glass containing milk is dropped or turned over, and what happens when a switch on the wall is pushed from one of its resting positions to another. While theories such as these are being formulated, a more general causal theory is also being formed. The more general theory posits that some events or states of nature are causes of specific effects. Without a general theory of causation, there would be no skeleton on which to hang the substance of the many specific causal theories that one needs to survive.

Nonetheless, the concepts of causation that are established early in life are too primitive to serve well as the basis for scientific theories. This shortcoming may be especially true in the health and social sciences, in which typical causes are neither necessary nor sufficient to bring about effects of interest. Hence, as has long been recognized in epidemiology, there is a need to develop a more refined conceptual model that can serve as a starting point in discussions of causation. In particular, such a model should address problems of multifactorial causation, confounding, interdependence of effects, direct and indirect effects, levels of causation, and systems or webs of causation (MacMahon and Pugh, 1967; Susser, 1973). This chapter describes one starting point, the sufficient-component cause model (or sufficient-cause model), which has proven useful in elucidating certain concepts in individual mechanisms of causation. Chapter 4 introduces the widely used potential-outcome or counterfactual model of causation, which is useful for relating individual-level to population-level causation, whereas Chapter 12 introduces graphical causal models (causal diagrams), which are especially useful for modeling causal systems.

Except where specified otherwise (in particular, in Chapter 27, on infectious disease), throughout the book we will assume that disease refers to a nonrecurrent event, such as death or first occurrence of a disease, and that the outcome of each individual or unit of study (e.g., a group of persons) is not affected by the exposures and outcomes of other individuals or units. Although this assumption will greatly simplify our discussion and is reasonable in many applications, it does not apply to contagious phenomena, such as transmissible behaviors and diseases. Nonetheless, all the definitions and most of the points we make (especially regarding validity) apply more generally. It is also essential to understand simpler situations before tackling the complexities created by causal interdependence of individuals or units.

A MODEL OF SUFFICIENT CAUSE AND COMPONENT CAUSES

To begin, we need to define *cause.* One definition of the cause of a specific disease occurrence is an antecedent event, condition, or characteristic that was necessary for the occurrence of the disease at the moment it occurred, given that other conditions are fixed. In other words, a cause of a disease occurrence is an event, condition, or characteristic that preceded the disease onset and that, had the event, condition, or characteristic been different in a specified way, the disease either would not have occurred at all or would not have occurred until some later time. Under this definition, if someone walking along an icy path falls and breaks a hip, there may be a long list of causes. These causes might include the weather on the day of the incident, the fact that the path was not cleared for pedestrians, the choice of footgear for the victim, the lack of a handrail, and so forth. The constellation of causes required for this particular person to break her hip at this particular time can be depicted with the sufficient cause diagrammed in Figure 2–1. By *sufficient cause* we mean a complete causal mechanism, a minimal set of conditions and events that are sufficient for the outcome to occur. The circle in the figure comprises five segments, each of which represents a causal component that must be present or have occured in order for the person to break her hip at that instant. The first component, labeled A, represents poor weather. The second component, labeled B, represents an uncleared path for pedestrians. The third component, labeled C, represents a poor choice of footgear. The fourth component, labeled D, represents the lack of a handrail. The final component, labeled U, represents all of the other unspecified events, conditions, and characteristics that must be present or have occured at the instance of the fall that led to a broken hip. For etiologic effects such as the causation of disease, many and possibly all of the components of a sufficient cause may be unknown (Rothman, 1976a). We usually include one component cause, labeled U, to represent the set of unknown factors.

All of the component causes in the sufficient cause are required and must be present or have occured at the instance of the fall for the person to break a hip. None is superfluous, which means that blocking the contribution of any component cause prevents the sufficient cause from acting. For many people, early causal thinking persists in attempts to find single causes as explanations for observed phenomena. But experience and reasoning show that the causal mechanism for any effect must consist of a constellation of components that act in concert (Mill, 1862; Mackie, 1965). In disease etiology, a sufficient cause is a set of conditions sufficient to ensure that the outcome will occur. Therefore, completing a sufficient cause is tantamount to the onset of disease. Onset here may refer to the onset of the earliest stage of the disease process or to any transition from one well-defined and readily characterized stage to the next, such as the onset of signs or symptoms.

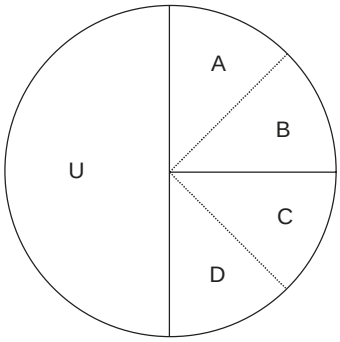


FIGURE 2–1 • Depiction of the constellation of component causes that constitute a sufficient cause for hip fracture for a particular person at a particular time. In the diagram, A represents poor weather, B represents an uncleared path for pedestrians, C represents a poor choice of footgear, D represents the lack of a handrail, and U represents all of the other unspecified events, conditions, and characteristics that must be present or must have occured at the instance of the fall that led to a broken hip.

Consider again the role of the handrail in causing hip fracture. The absence of such a handrail may play a causal role in some sufficient causes but not in others, depending on circumstances such as the weather, the level of inebriation of the pedestrian, and countless other factors. Our definition links the lack of a handrail with this one broken hip and does not imply that the lack of this handrail by itself was sufficient for that hip fracture to occur. With this definition of cause, no specific event, condition, or characteristic is sufficient by itself to produce disease. The definition does not describe a complete causal mechanism, but only a component of it. To say that the absence of a handrail is a component cause of a broken hip does not, however, imply that every person walking down the path will break a hip. Nor does it imply that if a handrail is installed with properties sufficient to prevent that broken hip, that no one will break a hip on that same path. There may be other sufficient causes by which a person could suffer a hip fracture. Each such sufficient cause would be depicted by its own diagram similar to Figure 2–1. The first of these sufficient causes to be completed by simultaneous accumulation of all of its component causes will be the one that depicts the mechanism by which the hip fracture occurs for a particular person. If no sufficient cause is completed while a person passes along the path, then no hip fracture will occur over the course of that walk.

As noted above, a characteristic of the naive concept of causation is the assumption of a one-to-one correspondence between the observed cause and effect. Under this view, each cause is seen as "necessary" and "sufficient" in itself to produce the effect, particularly when the cause is an observable action or event that takes place near in time to the effect. Thus, the flick of a switch appears to be the singular cause that makes an electric light go on. There are less evident causes, however, that also operate to produce the effect: a working bulb in the light fixture, intact wiring from the switch to the bulb, and voltage to produce a current when the circuit is closed. To achieve the effect of turning on the light, each of these components is as important as moving the switch, because changing any of these components of the causal constellation will prevent the effect. The term *necessary cause* is therefore reserved for a particular type of component cause under the sufficient-cause model. If any of the component causes appears in every sufficient cause, then that component cause is called a "necessary" component cause. For the disease to occur, any and all necessary component causes must be present or must have occurred. For example, one could label a component cause with the requirement that one must have a hip to suffer a hip fracture. Every sufficient cause that leads to hip fracture must have that component cause present, because in order to fracture a hip, one must have a hip to fracture.

The concept of complementary component causes will be useful in applications to epidemiology that follow. For each component cause in a sufficient cause, the set of the other component causes in that sufficient cause comprises the complementary component causes. For example, in Figure 2–1, component cause A (poor weather) has as its complementary component causes the components labeled B, C, D, and U. Component cause B (an uncleared path for pedestrians) has as its complementary component causes the components labeled A, C, D, and U.

THE NEED FOR A SPECIFIC REFERENCE CONDITION

Component causes must be defined with respect to a clearly specified alternative or reference condition (often called a *referent*). Consider again the lack of a handrail along the path. To say that this condition is a component cause of the broken hip, we have to specify an alternative condition against which to contrast the cause. The mere presence of a handrail would not suffice. After all, the hip fracture might still have occurred in the presence of a handrail, if the handrail was too short or if it was old and made of rotten wood. We might need to specify the presence of a handrail sufficiently tall and sturdy to break the fall for the absence of that handrail to be a component cause of the broken hip.

To see the necessity of specifying the alternative event, condition, or characteristic as well as the causal one, consider an example of a man who took high doses of ibuprofen for several years and developed a gastric ulcer. Did the man's use of ibuprofen cause his ulcer? One might at first assume that the natural contrast would be with what would have happened had he taken nothing instead of ibuprofen. Given a strong reason to take the ibuprofen, however, that alternative may not make sense. If the specified alternative to taking ibuprofen is to take acetaminophen, a different drug that might have been indicated for his problem, and if he would not have developed the ulcer had he used acetaminophen, then we can say that using ibuprofen caused the ulcer. But ibuprofen did not cause

his ulcer if the specified alternative is taking aspirin and, had he taken aspirin, he still would have developed the ulcer. The need to specify the alternative to a preventive is illustrated by a newspaper headline that read: "Rare Meat Cuts Colon Cancer Risk." Was this a story of an epidemiologic study comparing the colon cancer rate of a group of people who ate rare red meat with the rate in a group of vegetarians? No, the study compared persons who ate rare red meat with persons who ate highly cooked red meat. The same exposure, regular consumption of rare red meat, might have a preventive effect when contrasted against highly cooked red meat and a causative effect or no effect in contrast to a vegetarian diet. An event, condition, or characteristic is not a cause by itself as an intrinsic property it possesses in isolation, but as part of a causal contrast with an alternative event, condition, or characteristic (Lewis, 1973; Rubin, 1974; Greenland et al., 1999a; Maldonado and Greenland, 2002; see Chapter 4).

APPLICATION OF THE SUFFICIENT-CAUSE MODEL TO EPIDEMIOLOGY

The preceding introduction to concepts of sufficient causes and component causes provides the lexicon for application of the model to epidemiology. For example, tobacco smoking is a cause of lung cancer, but by itself it is not a sufficient cause, as demonstrated by the fact that most smokers do not get lung cancer. First, the term *smoking* is too imprecise to be useful beyond casual description. One must specify the type of smoke (e.g., cigarette, cigar, pipe, or environmental), whether it is filtered or unfiltered, the manner and frequency of inhalation, the age at initiation of smoking, and the duration of smoking. And, however smoking is defined, its alternative needs to be defined as well. Is it smoking nothing at all, smoking less, smoking something else? Equally important, even if smoking and its alternative are both defined explicitly, smoking will not cause cancer in everyone. So who is susceptible to this smoking effect? Or, to put it in other terms, what are the other components of the causal constellation that act with smoking to produce lung cancer in this contrast?

Figure 2–2 provides a schematic diagram of three sufficient causes that could be completed during the follow-up of an individual. The three conditions or events—A, B, and E—have been defined as binary variables, so they can only take on values of 0 or 1. With the coding of A used in the figure, its reference level, $A = 0$, is sometimes causative, but its index level, $A = 1$, is never causative. This situation arises because two sufficient causes contain a component cause labeled "$A = 0$," but no sufficient cause contains a component cause labeled "$A = 1$." An example of a condition or event of this sort might be $A = 1$ for taking a daily multivitamin supplement and $A = 0$ for taking no vitamin supplement. With the coding of B and E used in the example depicted by Figure 2–2, their index levels, $B = 1$ and $E = 1$, are sometimes causative, but their reference levels, $B = 0$ and $C = 0$, are never causative. For each variable, the index and reference levels may represent only two alternative states or events out of many possibilities. Thus, the coding of B might be $B = 1$ for smoking 20 cigarettes per day for 40 years and $B = 0$ for smoking 20 cigarettes per day for 20 years, followed by 20 years of not smoking. E might be coded $E = 1$ for living in an urban neighborhood with low average income and high income inequality, and $E = 0$ for living in an urban neighborhood with high average income and low income inequality.

$A = 0$, $B = 1$, and $E = 1$ are individual component causes of the sufficient causes in Figure 2–2. U_1, U_2, and U_3 represent sets of component causes. U_1, for example, is the set of all components other than $A = 0$ and $B = 1$ required to complete the first sufficient cause in Figure 2–2. If we decided not to specify $B = 1$, then $B = 1$ would become part of the set of components that are causally complementary to $A = 0$; in other words, $B = 1$ would then be absorbed into U_1.

Each of the three sufficient causes represented in Figure 2–2 is minimally sufficient to produce the disease in the individual. That is, only one of these mechanisms needs to be completed for

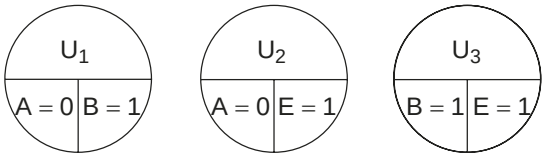


FIGURE 2–2 • Three classes of sufficient causes of a disease (sufficient causes I, II, and III from left to right).

disease to occur (sufficiency), and there is no superfluous component cause in any mechanism (minimality)—each component is a required part of that specific causal mechanism. A specific component cause may play a role in one, several, or all of the causal mechanisms. As noted earlier, a component cause that appears in all sufficient causes is called a *necessary* cause of the outcome. As an example, infection with HIV is a component of every sufficient cause of acquired immune deficiency syndrome (AIDS) and hence is a necessary cause of AIDS. It has been suggested that such causes be called "universally necessary," in recognition that every component of a sufficient cause is necessary for that sufficient cause (mechanism) to operate (Poole 2001a).

Figure 2–2 does not depict aspects of the causal process such as sequence or timing of action of the component causes, dose, or other complexities. These can be specified in the description of the contrast of index and reference conditions that defines each component cause. Thus, if the outcome is lung cancer and the factor B represents cigarette smoking, it might be defined more explicitly as smoking at least 20 cigarettes a day of unfiltered cigarettes for at least 40 years beginning at age 20 years or earlier (B = 1), or smoking 20 cigarettes a day of unfiltered cigarettes, beginning at age 20 years or earlier, and then smoking no cigarettes for the next 20 years (B = 0).

In specifying a component cause, the two sides of the causal contrast of which it is composed should be defined with an eye to realistic choices or options. If prescribing a placebo is not a realistic therapeutic option, a causal contrast between a new treatment and a placebo in a clinical trial may be questioned for its dubious relevance to medical practice. In a similar fashion, before saying that oral contraceptives increase the risk of death over 10 years (e.g., through myocardial infarction or stroke), we must consider the alternative to taking oral contraceptives. If it involves getting pregnant, then the risk of death attendant to childbirth might be greater than the risk from oral contraceptives, making oral contraceptives a preventive rather than a cause. If the alternative is an equally effective contraceptive without serious side effects, then oral contraceptives may be described as a cause of death.

To understand prevention in the sufficient-component cause framework, we posit that the alternative condition (in which a component cause is absent) prevents the outcome relative to the presence of the component cause. Thus, a preventive effect of a factor is represented by specifying its causative alternative as a component cause. An example is the presence of A = 0 as a component cause in the first two sufficient causes shown in Figure 2–2. Another example would be to define a variable, F (not depicted in Fig. 2–2), as "vaccination (F = 1) or no vaccination (F = 0)". Prevention of the disease by getting vaccinated (F = 1) would be expressed in the sufficient-component cause model as causation of the disease by not getting vaccinated (F = 0). This depiction is unproblematic because, once both sides of a causal contrast have been specified, causation and prevention are merely two sides of the same coin.

Sheps (1958) once asked, "Shall we count the living or the dead?" Death is an event, but survival is not. Hence, to use the sufficient-component cause model, we must count the dead. This model restriction can have substantive implications. For instance, some measures and formulas approximate others only when the outcome is rare. When survival is rare, death is common. In that case, use of the sufficient-component cause model to inform the analysis will prevent us from taking advantage of the rare-outcome approximations.

Similarly, etiologies of adverse health outcomes that are conditions or states, but not events, must be depicted under the sufficient-cause model by reversing the coding of the outcome. Consider spina bifida, which is the failure of the neural tube to close fully during gestation. There is no point in time at which spina bifida may be said to have occurred. It would be awkward to define the "incidence time" of spina bifida as the gestational age at which complete neural tube closure ordinarily occurs. The sufficient-component cause model would be better suited in this case to defining the event of complete closure (no spina bifida) as the outcome and to view conditions, events, and characteristics that prevent this beneficial event as the causes of the adverse condition of spina bifida.

PROBABILITY, RISK, AND CAUSES

In everyday language, "risk" is often used as a synonym for probability. It is also commonly used as a synonym for "hazard," as in, "Living near a nuclear power plant is a risk you should avoid." Unfortunately, in epidemiologic parlance, even in the scholarly literature, "risk" is frequently used for many distinct concepts: rate, rate ratio, risk ratio, incidence odds, prevalence, etc. The more

specific, and therefore more useful, definition of *risk* is "probability of an event during a specified period of time."

The term *probability* has multiple meanings. One is that it is the relative frequency of an event. Another is that probability is the tendency, or propensity, of an entity to produce an event. A third meaning is that probability measures someone's degree of certainty that an event will occur. When one says "the probability of death in vehicular accidents when traveling >120 km/h is high," one means that the proportion of accidents that end with deaths is higher when they involve vehicles traveling >120 km/h than when they involve vehicles traveling at lower speeds (frequency usage), that high-speed accidents have a greater tendency than lower-speed accidents to result in deaths (propensity usage), or that the speaker is more certain that a death will occur in a high-speed accident than in a lower-speed accident (certainty usage).

The frequency usage of "probability" and "risk," unlike the propensity and certainty usages, admits no meaning to the notion of "risk" for an individual beyond the relative frequency of 100% if the event occurs and 0% if it does not. This restriction of individual risks to 0 or 1 can only be relaxed to allow values in between by reinterpreting such statements as the frequency with which the outcome would be seen upon random sampling from a very large population of individuals deemed to be "like" the individual in some way (e.g., of the same age, sex, and smoking history). If one accepts this interpretation, whether any actual sampling has been conducted or not, the notion of individual risk is replaced by the notion of the frequency of the event in question in the large population from which the individual was sampled. With this view of risk, a risk will change according to how we group individuals together to evaluate frequencies. Subjective judgment will inevitably enter into the picture in deciding which characteristics to use for grouping. For instance, should tomato consumption be taken into account in defining the class of men who are "like" a given man for purposes of determining his risk of a diagnosis of prostate cancer between his 60th and 70th birthdays? If so, which study or meta-analysis should be used to factor in this piece of information?

Unless we have found a set of conditions and events in which the disease does not occur at all, it is always a reasonable working hypothesis that, no matter how much is known about the etiology of a disease, some causal components remain unknown. We may be inclined to assign an equal risk to all individuals whose status for some components is known and identical. We may say, for example, that men who are heavy cigarette smokers have approximately a 10% lifetime risk of developing lung cancer. Some interpret this statement to mean that all men would be subject to a 10% probability of lung cancer if they were to become heavy smokers, as if the occurrence of lung cancer, aside from smoking, were purely a matter of chance. This view is untenable. A probability may be 10% conditional on one piece of information and higher or lower than 10% if we condition on other relevant information as well. For instance, men who are heavy cigarette smokers and who worked for many years in occupations with historically high levels of exposure to airborne asbestos fibers would be said to have a lifetime lung cancer risk appreciably higher than 10%.

Regardless of whether we interpret probability as relative frequency or degree of certainty, the assignment of equal risks merely reflects the particular grouping. In our ignorance, the best we can do in assessing risk is to classify people according to measured risk indicators and then assign the average risk observed within a class to persons within the class. As knowledge or specification of additional risk indicators expands, the risk estimates assigned to people will depart from average according to the presence or absence of other factors that predict the outcome.

STRENGTH OF EFFECTS

The causal model exemplified by Figure 2–2 can facilitate an understanding of some key concepts such as *strength of effect* and *interaction.* As an illustration of strength of effect, Table 2–1 displays the frequency of the eight possible patterns for exposure to A, B, and E in two hypothetical populations. Now the pie charts in Figure 2–2 depict classes of mechanisms. The first one, for instance, represents all sufficient causes that, no matter what other component causes they may contain, have in common the fact that they contain $A = 0$ and $B = 1$. The constituents of U_1 may, and ordinarily would, differ from individual to individual. For simplification, we shall suppose, rather unrealistically, that U_1, U_2, and U_3 are always present or have always occured for everyone and Figure 2–2 represents all the sufficient causes.

TABLE 2–1

Exposure Frequencies and Individual Risks in Two Hypothetical Populations According to the Possible Combinations of the Three Specified Component Causes in Fig. 2–1

Exposures					Frequency of Exposure Pattern	
A	B	E	Sufficient Cause Completed	Risk	Population 1	Population 2
1	1	1	III	1	900	100
1	1	0	None	0	900	100
1	0	1	None	0	100	900
1	0	0	None	0	100	900
0	1	1	I, II, or III	1	100	900
0	1	0	I	1	100	900
0	0	1	II	1	900	100
0	0	0	none	0	900	100

Under these assumptions, the response of each individual to the exposure pattern in a given row can be found in the response column. The response here is the risk of developing a disease over a specified time period that is the same for all individuals. For simplification, a deterministic model of risk is employed, such that individual risks can equal only the value 0 or 1, and no values in between. A stochastic model of individual risk would relax this restriction and allow individual risks to lie between 0 and 1.

The proportion getting disease, or incidence proportion, in any subpopulation in Table 2–1 can be found by summing the number of persons at each exposure pattern with an individual risk of 1 and dividing this total by the subpopulation size. For example, if exposure A is not considered (e.g., if it were not measured), the pattern of incidence proportions in population 1 would be those in Table 2–2.

As an example of how the proportions in Table 2–2 were calculated, let us review how the incidence proportion among persons in population 1 with $B = 1$ and $E = 0$ was calculated: There were 900 persons with $A = 1$, $B = 1$, and $E = 0$, none of whom became cases because there are no sufficient causes that can culminate in the occurrence of the disease over the study period in persons with this combination of exposure conditions. (There are two sufficient causes that contain $B = 1$ as a component cause, but one of them contains the component cause $A = 0$ and the other contains the component cause $E = 1$. The presence of $A = 1$ or $E = 0$ blocks these etiologic mechanisms.) There were 100 persons with $A = 0$, $B = 1$, and $E = 0$, all of whom became cases because they all had U_1, the set of causal complements for the class of sufficient causes containing $A = 0$ and

TABLE 2–2

Incidence Proportions (IP) for Combinations of Component Causes B and E in Hypothetical Population 1, Assuming That Component Cause A Is Unmeasured

	B = 1, E = 1	B = 1, E = 0	B = 0, E = 1	B = 0, E = 0
Cases	1,000	100	900	0
Total	1,000	1,000	1,000	1,000
IP	1.00	0.10	0.90	0.00

TABLE 2–3

Incidence Proportions (IP) for Combinations of Component Causes B and E in Hypothetical Population 2, Assuming That Component Cause A Is Unmeasured

	B = 1, E = 1	B = 1, E = 0	B = 0, E = 1	B = 0, E = 0
Cases	1,000	900	100	0
Total	1,000	1,000	1,000	1,000
IP	1.00	0.90	0.10	0.00

B = 1. Thus, among all 1,000 persons with B = 1 and E = 0, there were 100 cases, for an incidence proportion of 0.10.

If we were to measure strength of effect by the difference of the incidence proportions, it is evident from Table 2–2 that for population 1, E = 1 has a much stronger effect than B = 1, because E = 1 increases the incidence proportion by 0.9 (in both levels of B), whereas B = 1 increases the incidence proportion by only 0.1 (in both levels of E). Table 2–3 shows the analogous results for population 2. Although the members of this population have exactly the same causal mechanisms operating within them as do the members of population 1, the relative strengths of causative factors E = 1 and B = 1 are reversed, again using the incidence proportion difference as the measure of strength. B = 1 now has a much stronger effect on the incidence proportion than E = 1, despite the fact that A, B, and E have no association with one another in either population, and their index levels (A = 1, B = 1 and E = 1) and reference levels (A = 0, B = 0, and E = 0) are each present or have occured in exactly half of each population.

The overall difference of incidence proportions contrasting E = 1 with E = 0 is (1,900/2,000) − (100/2,000) = 0.9 in population 1 and (1,100/2,000) − (900/2,000) = 0.1 in population 2. The key difference between populations 1 and 2 is the difference in the prevalence of the conditions under which E = 1 acts to increase risk: that is, the presence of A = 0 or B = 1, but not both. (When A = 0 *and* B = 1, E = 1 completes all three sufficient causes in Figure 2–2; it thus does not increase anyone's risk, although it may well shorten the time to the outcome.) The prevalence of the condition, "A = 0 or B = 1 but not both" is 1,800/2,000 = 90% in both levels of E in population 1. In population 2, this prevalence is only 200/2,000 = 10% in both levels of E. This difference in the prevalence of the conditions sufficient for E = 1 to increase risk explains the difference in the strength of the effect of E = 1 as measured by the difference in incidence proportions.

As noted above, the set of all other component causes in all sufficient causes in which a causal factor participates is called the *causal complement* of the factor. Thus, A = 0, B = 1, U_2, and U_3 make up the causal complement of E = 1 in the above example. This example shows that the strength of a factor's effect on the occurrence of a disease in a population, measured as the absolute difference in incidence proportions, depends on the prevalence of its causal complement. This dependence has nothing to do with the etiologic mechanism of the component's action, because the component is an equal partner in each mechanism in which it appears. Nevertheless, a factor will appear to have a strong effect, as measured by the difference of proportions getting disease, if its causal complement is common. Conversely, a factor with a rare causal complement will appear to have a weak effect.

If strength of effect is measured by the ratio of proportions getting disease, as opposed to the difference, then strength depends on more than a factor's causal complement. In particular, it depends additionally on how common or rare the components are of sufficient causes in which the specified causal factor does *not* play a role. In this example, given the ubiquity of U_1, the effect of E = 1 measured in ratio terms depends on the prevalence of E = 1's causal complement and on the prevalence of the conjunction of A = 0 and B = 1. If many people have both A = 0 and B = 1, the "baseline" incidence proportion (i.e., the proportion of not-E or "unexposed" persons getting disease) will be high and the proportion getting disease due to E will be comparatively low. If few

people have both A = 0 and B = 1, the baseline incidence proportion will be low and the proportion getting disease due to E = 1 will be comparatively high. Thus, strength of effect measured by the incidence proportion ratio depends on more conditions than does strength of effect measured by the incidence proportion difference.

Regardless of how strength of a causal factor's effect is measured, the public health significance of that effect does not imply a corresponding degree of etiologic significance. Each component cause in a given sufficient cause has the same etiologic significance. Given a specific causal mechanism, any of the component causes can have strong or weak effects using either the difference or ratio measure. The actual identities of the components of a sufficient cause are part of the mechanics of causation, whereas the strength of a factor's effect depends on the time-specific distribution of its causal complement (if strength is measured in absolute terms) plus the distribution of the components of all sufficient causes in which the factor does not play a role (if strength is measured in relative terms). Over a span of time, the strength of the effect of a given factor on disease occurrence may change because the prevalence of its causal complement in various mechanisms may also change, even if the causal mechanisms in which the factor and its cofactors act remain unchanged.

INTERACTION AMONG CAUSES

Two component causes acting in the same sufficient cause may be defined as *interacting causally* to produce disease. This definition leaves open many possible mechanisms for the interaction, including those in which two components interact in a direct physical fashion (e.g., two drugs that react to form a toxic by-product) and those in which one component (the *initiator* of the pair) alters a substrate so that the other component (the *promoter* of the pair) can act. Nonetheless, it excludes any situation in which one component E is merely a cause of another component F, with no effect of E on disease except through the component F it causes.

Acting in the same sufficient cause is not the same as one component cause acting to produce a second component cause, and then the second component going on to produce the disease (Robins and Greenland 1992, Kaufman et al., 2004). As an example of the distinction, if cigarette smoking (vs. never smoking) is a component cause of atherosclerosis, and atherosclerosis (vs. no atherosclerosis) causes myocardial infarction, both smoking and atherosclerosis would be component causes (cofactors) in certain sufficient causes of myocardial infarction. They would not necessarily appear in the same sufficient cause. Rather, for a sufficient cause involving atherosclerosis as a component cause, there would be another sufficient cause in which the atherosclerosis component cause was replaced by all the component causes that brought about the atherosclerosis, including smoking. Thus, a sequential causal relation between smoking and atherosclerosis would not be enough for them to interact synergistically in the etiology of myocardial infarction, in the sufficient-cause sense. Instead, the causal sequence means that smoking can act indirectly, through atherosclerosis, to bring about myocardial infarction.

Now suppose that, perhaps in addition to the above mechanism, smoking reduces clotting time and thus causes thrombi that block the coronary arteries if they are narrowed by atherosclerosis. This mechanism would be represented by a sufficient cause containing both smoking and atherosclerosis as components and thus would constitute a synergistic interaction between smoking and atherosclerosis in causing myocardial infarction. The presence of this sufficient cause would not, however, tell us whether smoking also contributed to the myocardial infarction by causing the atherosclerosis. Thus, the basic sufficient-cause model does not alert us to indirect effects (effects of some component causes mediated by other component causes in the model). Chapters 4 and 12 introduce potential-outcome and graphical models better suited to displaying indirect effects and more general sequential mechanisms, whereas Chapter 5 discusses in detail interaction as defined in the potential-outcome framework and its relation to interaction as defined in the sufficient-cause model.

PROPORTION OF DISEASE DUE TO SPECIFIC CAUSES

In Figure 2–2, assuming that the three sufficient causes in the diagram are the only ones operating, what fraction of disease is caused by E = 1? E = 1 is a component cause of disease in two of the sufficient-cause mechanisms, II and III, so all disease arising through either of these two mechanisms is attributable to E = 1. Note that in persons with the exposure pattern A = 0, B = 1, E = 1, all three

sufficient causes would be completed. The first of the three mechanisms to be completed would be the one that actually produces a given case. If the first one completed is mechanism II or III, the case would be causally attributable to E = 1. If mechanism I is the first one to be completed, however, E = 1 would not be part of the sufficient cause producing that case. Without knowing the completion times of the three mechanisms, among persons with the exposure pattern A = 0, B = 1, E = 1 we cannot tell how many of the 100 cases in population 1 or the 900 cases in population 2 are etiologically attributable to E = 1.

Each of the cases that is etiologically attributable to E = 1 can also be attributed to the other component causes in the causal mechanisms in which E = 1 acts. Each component cause interacts with its complementary factors to produce disease, so each case of disease can be attributed to every component cause in the completed sufficient cause. Note, though, that the attributable fractions added across component causes of the same disease do not sum to 1, although there is a mistaken tendency to think that they do. To illustrate the mistake in this tendency, note that a necessary component cause appears in every completed sufficient cause of disease, and so by itself has an attributable fraction of 1, without counting the attributable fractions for other component causes. Because every case of disease can be attributed to every component cause in its causal mechanism, attributable fractions for different component causes will generally sum to more than 1, and there is no upper limit for this sum.

A recent debate regarding the proportion of risk factors for coronary heart disease attributable to particular component causes illustrates the type of errors in inference that can arise when the sum is thought to be restricted to 1. The debate centers around whether the proportion of coronary heart disease attributable to high blood cholesterol, high blood pressure, and cigarette smoking equals 75% or "only 50%" (Magnus and Beaglehole, 2001). If the former, then some have argued that the search for additional causes would be of limited utility (Beaglehole and Magnus, 2002), because only 25% of cases "remain to be explained." By assuming that the proportion explained by yet unknown component causes cannot exceed 25%, those who support this contention fail to recognize that cases caused by a sufficient cause that contains any subset of the three named causes might also contain unknown component causes. Cases stemming from sufficient causes with this overlapping set of component causes could be prevented by interventions targeting the three named causes, or by interventions targeting the yet unknown causes when they become known. The latter interventions could reduce the disease burden by much more than 25%.

As another example, in a cohort of cigarette smokers exposed to arsenic by working in a smelter, an estimated 75% of the lung cancer rate was attributable to their work environment and an estimated 65% was attributable to their smoking (Pinto et al., 1978; Hertz-Picciotto et al., 1992). There is no problem with such figures, which merely reflect the multifactorial etiology of disease. So, too, with coronary heart disease; if 75% of that disease is attributable to high blood cholesterol, high blood pressure, and cigarette smoking, 100% of it can still be attributable to other causes, known, suspected, and yet to be discovered. Some of these causes will participate in the same causal mechanisms as high blood cholesterol, high blood pressure, and cigarette smoking. Beaglehole and Magnus were correct in thinking that if the three specified component causes combine to explain 75% of cardiovascular disease (CVD) and we somehow eliminated them, there would be only 25% of CVD cases remaining. But until that 75% is eliminated, any newly discovered component could cause up to 100% of the CVD we currently have.

The notion that interventions targeting high blood cholesterol, high blood pressure, and cigarette smoking could eliminate 75% of coronary heart disease is unrealistic given currently available intervention strategies. Although progress can be made to reduce the effect of these risk factors, it is unlikely that any of them could be completely eradicated from any large population in the near term. Estimates of the public health effect of eliminating diseases themselves as causes of death (Murray et al., 2002) are even further removed from reality, because they fail to account for all the effects of interventions required to achieve the disease elimination, including unanticipated side effects (Greenland, 2002a, 2005a).

The debate about coronary heart disease attribution to component causes is reminiscent of an earlier debate regarding causes of cancer. In their widely cited work, *The Causes of Cancer*, Doll and Peto (1981, Table 20) created a table giving their estimates of the fraction of all cancers caused by various agents. The fractions summed to nearly 100%. Although the authors acknowledged that any case could be caused by more than one agent (which means that, given enough agents, the attributable

fractions would sum to far more than 100%), they referred to this situation as a "difficulty" and an "anomaly" that they chose to ignore. Subsequently, one of the authors acknowledged that the attributable fraction could sum to greater than 100% (Peto, 1985). It is neither a difficulty nor an anomaly nor something we can safely ignore, but simply a consequence of the fact that no event has a single agent as the cause. The fraction of disease that can be attributed to known causes will grow without bound as more causes are discovered. Only the fraction of disease attributable to a single component cause cannot exceed 100%.

In a similar vein, much publicity attended the pronouncement in 1960 that as much as 90% of cancer is environmentally caused (Higginson, 1960). Here, "environment" was thought of as representing all nongenetic component causes, and thus included not only the physical environment, but also the social environment and all individual human behavior that is not genetically determined. Hence, environmental component causes must be present to some extent in every sufficient cause of a disease. Thus, Higginson's estimate of 90% was an underestimate.

One can also show that 100% of any disease is inherited, even when environmental factors are component causes. MacMahon (1968) cited the example given by Hogben (1933) of yellow shanks, a trait occurring in certain genetic strains of fowl fed on yellow corn. Both a particular set of genes and a yellow-corn diet are necessary to produce yellow shanks. A farmer with several strains of fowl who feeds them all only yellow corn would consider yellow shanks to be a genetic condition, because only one strain would get yellow shanks, despite all strains getting the same diet. A different farmer who owned only the strain liable to get yellow shanks but who fed some of the birds yellow corn and others white corn would consider yellow shanks to be an environmentally determined condition because it depends on diet. In humans, the mental retardation caused by phenylketonuria is considered by many to be purely genetic. This retardation can, however, be successfully prevented by dietary intervention, which demonstrates the presence of an environmental cause. In reality, yellow shanks, phenylketonuria, and other diseases and conditions are determined by an interaction of genes and environment. It makes no sense to allocate a portion of the causation to either genes or environment separately when both may act together in sufficient causes.

Nonetheless, many researchers have compared disease occurrence in identical and nonidentical twins to estimate the fraction of disease that is inherited. These twin-study and other heritability indices assess only the relative role of environmental and genetic causes of disease in a particular setting. For example, some genetic causes may be necessary components of every causal mechanism. If everyone in a population has an identical set of the genes that cause disease, however, their effect is not included in heritability indices, despite the fact that the genes are causes of the disease. The two farmers in the preceding example would offer very different values for the heritability of yellow shanks, despite the fact that the condition is always 100% dependent on having certain genes.

Every case of every disease has some environmental and some genetic component causes, and therefore every case can be attributed both to genes and to environment. No paradox exists as long as it is understood that the fractions of disease attributable to genes and to environment overlap with one another. Thus, debates over what proportion of all occurrences of a disease are genetic and what proportion are environmental, inasmuch as these debates assume that the shares must add up to 100%, are fallacious and distracting from more worthwhile pursuits.

On an even more general level, the question of whether a given disease does or does not have a "multifactorial etiology" can be answered once and for all in the affirmative. All diseases have multifactorial etiologies. It is therefore completely unremarkable for a given disease to have such an etiology, and no time or money should be spent on research trying to answer the question of whether a particular disease does or does not have a multifactorial etiology. They all do. The job of etiologic research is to identify components of those etiologies.

INDUCTION PERIOD

Pie-chart diagrams of sufficient causes and their components such as those in Figure 2–2 are not well suited to provide a model for conceptualizing the *induction period*, which may be defined as the period of time from causal action until disease initiation. There is no way to tell from a pie-chart diagram of a sufficient cause which components affect each other, which components must come before or after others, for which components the temporal order is irrelevant, etc. The crucial

information on temporal ordering must come in a separate description of the interrelations among the components of a sufficient cause.

If, in sufficient cause I, the sequence of action of the specified component causes must be A = 0, B = 1 and we are studying the effect of A = 0, which (let us assume) acts at a narrowly defined point in time, we do not observe the occurrence of disease immediately after A = 0 occurs. Disease occurs only after the sequence is completed, so there will be a delay while B = 1 occurs (along with components of the set U_1 that are not present or that have not occured when A = 0 occurs). When B = 1 acts, if it is the last of all the component causes (including those in the set of unspecified conditions and events represented by U_1), disease occurs. The interval between the action of B = 1 and the disease occurrence is the induction time for the effect of B = 1 in sufficient cause I.

In the example given earlier of an equilibrium disorder leading to a later fall and hip injury, the induction time between the start of the equilibrium disorder and the later hip injury might be long, if the equilibrium disorder is caused by an old head injury, or short, if the disorder is caused by inebriation. In the latter case, it could even be instantaneous, if we define it as blood alcohol greater than a certain level. This latter possibility illustrates an important general point: Component causes that do not change with time, as opposed to events, all have induction times of zero.

Defining an induction period of interest is tantamount to specifying the characteristics of the component causes of interest. A clear example of a lengthy induction time is the cause–effect relation between exposure of a female fetus to diethylstilbestrol (DES) and the subsequent development of adenocarcinoma of the vagina. The cancer is usually diagnosed between ages 15 and 30 years. Because the causal exposure to DES occurs early in pregnancy, there is an induction time of about 15 to 30 years for the carcinogenic action of DES. During this time, other causes presumably are operating; some evidence suggests that hormonal action during adolescence may be part of the mechanism (Rothman, 1981).

It is incorrect to characterize a disease itself as having a lengthy or brief induction period. The induction time can be conceptualized only in relation to a specific component cause operating in a specific sufficient cause. Thus, we say that the induction time relating DES to clear-cell carcinoma of the vagina is 15 to 30 years, but we should not say that 15 to 30 years is the induction time for clear-cell carcinoma in general. Because each component cause in any causal mechanism can act at a time different from the other component causes, each can have its own induction time. For the component cause that acts last, the induction time equals zero. If another component cause of clear-cell carcinoma of the vagina that acts during adolescence were identified, it would have a much shorter induction time for its carcinogenic action than DES. Thus, induction time characterizes a specific cause–effect pair rather than just the effect.

In carcinogenesis, the terms *initiator* and *promotor* have been used to refer to some of the component causes of cancer that act early and late, respectively, in the causal mechanism. Cancer itself has often been characterized as a disease process with a long induction time. This characterization is a misconception, however, because any late-acting component in the causal process, such as a promotor, will have a short induction time. Indeed, by definition, the induction time will always be zero for at least one component cause, the last to act. The mistaken view that diseases, as opposed to cause–disease relationships, have long or short induction periods can have important implications for research. For instance, the view of adult cancers as "diseases of long latency" may induce some researchers to ignore evidence of etiologic effects occurring relatively late in the processes that culminate in clinically diagnosed cancers. At the other extreme, the routine disregard for exposures occurring in the first decade or two in studies of occupational carcinogenesis, as a major example, may well have inhibited the discovery of occupational causes with very long induction periods.

Disease, once initiated, will not necessarily be apparent. The time interval between irreversible disease occurrence and detection has been termed the *latent period* (Rothman, 1981), although others have used this term interchangeably with induction period. Still others use *latent period* to mean the total time between causal action and disease detection. We use *induction period* to describe the time from causal action to irreversible disease occurrence and *latent period* to mean the time from disease occurrence to disease detection. The latent period can sometimes be reduced by improved methods of disease detection. The induction period, on the other hand, cannot be reduced by early detection of disease, because disease occurrence marks the end of the induction period. Earlier detection of disease, however, may reduce the apparent induction period (the time between causal action and disease detection), because the time when disease is detected, as a practical matter, is

usually used to mark the time of disease occurrence. Thus, diseases such as slow-growing cancers may appear to have long induction periods with respect to many causes because they have long latent periods. The latent period, unlike the induction period, is a characteristic of the disease and the detection effort applied to the person with the disease.

Although it is not possible to reduce the induction period proper by earlier detection of disease, it may be possible to observe intermediate stages of a causal mechanism. The increased interest in biomarkers such as DNA adducts is an example of attempting to focus on causes more proximal to the disease occurrence or on effects more proximal to cause occurrence. Such biomarkers may nonetheless reflect the effects of earlier-acting agents on the person.

Some agents may have a causal action by shortening the induction time of other agents. Suppose that exposure to factor $X = 1$ leads to epilepsy after an interval of 10 years, on average. It may be that exposure to a drug, $Z = 1$, would shorten this interval to 2 years. Is $Z = 1$ acting as a catalyst, or as a cause, of epilepsy? The answer is both: A catalyst is a cause. Without $Z = 1$, the occurrence of epilepsy comes 8 years later than it comes with $Z = 1$, so we can say that $Z = 1$ causes the onset of the early epilepsy. It is not sufficient to argue that the epilepsy would have occurred anyway. First, it would not have occurred at that time, and the time of occurrence is part of our definition of an event. Second, epilepsy will occur later only if the individual survives an additional 8 years, which is not certain. Not only does agent $Z = 1$ determine when the epilepsy occurs, it can also determine whether it occurs. Thus, we should call any agent that acts as a catalyst of a causal mechanism, speeding up an induction period for other agents, a cause in its own right. Similarly, any agent that postpones the onset of an event, drawing out the induction period for another agent, is a preventive. It should not be too surprising to equate postponement to prevention: We routinely use such an equation when we employ the euphemism that we "prevent" death, which actually can only be postponed. What we prevent is death at a given time, in favor of death at a later time.

SCOPE OF THE MODEL

The main utility of this model of sufficient causes and their components lies in its ability to provide a general but practical conceptual framework for causal problems. The attempt to make the proportion of disease attributable to various component causes add to 100% is an example of a fallacy that is exposed by the model (although MacMahon and others were able to invoke yellow shanks and phenylketonuria to expose that fallacy long before the sufficient-component cause model was formally described [MacMahon and Pugh, 1967, 1970]). The model makes it clear that, because of interactions, there is no upper limit to the sum of these proportions. As we shall see in Chapter 5, the epidemiologic evaluation of interactions themselves can be clarified, to some extent, with the help of the model.

Although the model appears to deal qualitatively with the action of component causes, it can be extended to account for dose dependence by postulating a set of sufficient causes, each of which contains as a component a different dose of the agent in question. Small doses might require a larger or rarer set of complementary causes to complete a sufficient cause than that required by large doses (Rothman, 1976a), in which case it is particularly important to specify both sides of the causal contrast. In this way, the model can account for the phenomenon of a shorter induction period accompanying larger doses of exposure, because a smaller set of complementary components would be needed to complete the sufficient cause.

Those who believe that chance must play a role in any complex mechanism might object to the intricacy of this seemingly deterministic model. A probabilistic (stochastic) model could be invoked to describe a dose–response relation, for example, without the need for a multitude of different causal mechanisms. The model would simply relate the dose of the exposure to the probability of the effect occurring. For those who believe that virtually all events contain some element of chance, deterministic causal models may seem to misrepresent the indeterminism of the real world. However, the deterministic model presented here can accommodate "chance"; one way might be to view chance, or at least some part of the variability that we call "chance," as the result of deterministic events that are beyond the current limits of knowledge or observability.

For example, the outcome of a flip of a coin is usually considered a chance event. In classical mechanics, however, the outcome can in theory be determined completely by the application of physical laws and a sufficient description of the starting conditions. To put it in terms more familiar

to epidemiologists, consider the explanation for why an individual gets lung cancer. One hundred years ago, when little was known about the etiology of lung cancer; a scientist might have said that it was a matter of chance. Nowadays, we might say that the risk depends on how much the individual smokes, how much asbestos and radon the individual has been exposed to, and so on. Nonetheless, recognizing this dependence moves the line of ignorance; it does not eliminate it. One can still ask what determines whether an individual who has smoked a specific amount and has a specified amount of exposure to all the other known risk factors will get lung cancer. Some will get lung cancer and some will not, and if all known risk factors are already taken into account, what is left we might still describe as chance. True, we can explain much more of the variability in lung cancer occurrence nowadays than we formerly could by taking into account factors known to cause it, but at the limits of our knowledge, we still ascribe the remaining variability to what we call chance. In this view, chance is seen as a catchall term for our ignorance about causal explanations.

We have so far ignored more subtle considerations of sources of unpredictability in events, such as chaotic behavior (in which even the slightest uncertainty about initial conditions leads to vast uncertainty about outcomes) and quantum-mechanical uncertainty. In each of these situations, a random (stochastic) model component may be essential for any useful modeling effort. Such components can also be introduced in the above conceptual model by treating unmeasured component causes in the model as random events, so that the causal model based on components of sufficient causes can have random elements. An example is treatment assignment in randomized clinical trials (Poole 2001a).

OTHER MODELS OF CAUSATION

The sufficient-component cause model is only one of several models of causation that may be useful for gaining insight about epidemiologic concepts (Greenland and Brumback, 2002; Greenland, 2004a). It portrays qualitative causal mechanisms within members of a population, so its fundamental unit of analysis is the causal mechanism rather than a person. Many different sets of mechanisms can lead to the same pattern of disease within a population, so the sufficient-component cause model involves specification of details that are beyond the scope of epidemiologic data. Also, it does not incorporate elements reflecting population distributions of factors or causal sequences, which are crucial to understanding confounding and other biases.

Other models of causation, such as potential-outcome (counterfactual) models and graphical models, provide direct representations of epidemiologic concepts such as confounding and other biases, and can be applied at mechanistic, individual, or population levels of analysis. Potential-outcome models (Chapters 4 and 5) specify in detail what would happen to individuals or populations under alternative possible patterns of interventions or exposures, and also bring to the fore problems in operationally defining causes (Greenland, 2002a, 2005a; Hernán, 2005). Graphical models (Chapter 12) display broad qualitative assumptions about causal directions and independencies. Both types of model have close relationships to the structural-equations models that are popular in the social sciences (Pearl, 2000; Greenland and Brumback, 2002), and both can be subsumed under a general theory of longitudinal causality (Robins, 1997).

PHILOSOPHY OF SCIENTIFIC INFERENCE

Causal inference may be viewed as a special case of the more general process of scientific reasoning. The literature on this topic is too vast for us to review thoroughly, but we will provide a brief overview of certain points relevant to epidemiology, at the risk of some oversimplification.

INDUCTIVISM

Modern science began to emerge around the 16th and 17th centuries, when the knowledge demands of emerging technologies (such as artillery and transoceanic navigation) stimulated inquiry into the origins of knowledge. An early codification of the scientific method was Francis Bacon's *Novum Organum*, which, in 1620, presented an inductivist view of science. In this philosophy, scientific reasoning is said to depend on making generalizations, or inductions, from observations to general laws of nature; the observations are said to induce the formulation of a natural law in the mind of

the scientist. Thus, an inductivist would have said that Jenner's observation of lack of smallpox among milkmaids induced in Jenner's mind the theory that cowpox (common among milkmaids) conferred immunity to smallpox. Inductivist philosophy reached a pinnacle of sorts in the canons of John Stuart Mill (1862), which evolved into inferential criteria that are still in use today.

Inductivist philosophy was a great step forward from the medieval scholasticism that preceded it, for at least it demanded that a scientist make careful observations of people and nature rather than appeal to faith, ancient texts, or authorities. Nonetheless, in the 18th century the Scottish philosopher David Hume described a disturbing deficiency in inductivism. An inductive argument carried no logical force; instead, such an argument represented nothing more than an *assumption* that certain events would in the future follow the same pattern as they had in the past. Thus, to argue that cowpox caused immunity to smallpox because no one got smallpox after having cowpox corresponded to an unjustified assumption that the pattern observed to date (no smallpox after cowpox) would continue into the future. Hume pointed out that, even for the most reasonable-sounding of such assumptions, there was no logical necessity behind the inductive argument.

Of central concern to Hume (1739) was the issue of causal inference and failure of induction to provide a foundation for it:

> Thus not only our reason fails us in the discovery of the ultimate connexion of causes and effects, but even after experience has inform'd us of their constant conjunction, 'tis impossible for us to satisfy ourselves by our reason, why we shou'd extend that experience beyond those particular instances, which have fallen under our observation. We suppose, but are never able to prove, that there must be a resemblance betwixt those objects, of which we have had experience, and those which lie beyond the reach of our discovery.

In other words, no number of repetitions of a particular sequence of events, such as the appearance of a light after flipping a switch, can prove a causal connection between the action of the switch and the turning on of the light. No matter how many times the light comes on after the switch has been pressed, the possibility of coincidental occurrence cannot be ruled out. Hume pointed out that observers cannot perceive causal connections, but only a series of events. Bertrand Russell (1945) illustrated this point with the example of two accurate clocks that perpetually chime on the hour, with one keeping time slightly ahead of the other. Although one invariably chimes before the other, there is no direct causal connection from one to the other. Thus, assigning a causal interpretation to the pattern of events cannot be a logical extension of our observations alone, because the events might be occurring together only because of a shared earlier cause, or because of some systematic error in the observations.

Causal inference based on mere association of events constitutes a logical fallacy known as *post hoc ergo propter hoc* (Latin for "after this therefore on account of this"). This fallacy is exemplified by the inference that the crowing of a rooster is necessary for the sun to rise because sunrise is always preceded by the crowing.

The *post hoc* fallacy is a special case of a more general logical fallacy known as the *fallacy of affirming the consequent.* This fallacy of confirmation takes the following general form: "We know that if H is true, B must be true; and we know that B is true; therefore H must be true." This fallacy is used routinely by scientists in interpreting data. It is used, for example, when one argues as follows: "If sewer service causes heart disease, then heart disease rates should be highest where sewer service is available; heart disease rates are indeed highest where sewer service is available; therefore, sewer service causes heart disease." Here, H is the hypothesis "sewer service causes heart disease" and B is the observation "heart disease rates are highest where sewer service is available." The argument is logically unsound, as demonstrated by the fact that we can imagine many ways in which the premises could be true but the conclusion false; for example, economic development could lead to both sewer service and elevated heart disease rates, without any effect of sewer service on heart disease. In this case, however, we also know that one of the premises is not true—specifically, the premise, "If H is true, B must be true." This particular form of the fallacy exemplifies the problem of *confounding*, which we will discuss in detail in later chapters.

Bertrand Russell (1945) satirized the fallacy this way:

> 'If p, then q; now q is true; therefore p is true.' E.g., 'If pigs have wings, then some winged animals are good to eat; now some winged animals are good to eat; therefore pigs have wings.' This form of inference is called 'scientific method.'

REFUTATIONISM

Russell was not alone in his lament of the illogicality of scientific reasoning as ordinarily practiced. Many philosophers and scientists from Hume's time forward attempted to set out a firm logical basis for scientific reasoning.

In the 1920s, most notable among these was the school of logical positivists, who sought a logic for science that could lead inevitably to correct scientific conclusions, in much the way rigorous logic can lead inevitably to correct conclusions in mathematics. Other philosophers and scientists, however, had started to suspect that scientific hypotheses can never be proven or established as true in any logical sense. For example, a number of philosophers noted that scientific statements can only be found to be consistent with observation, but cannot be proven or disproven in any "airtight" logical or mathematical sense (Duhem, 1906, transl. 1954; Popper 1934, transl. 1959; Quine, 1951). This fact is sometimes called the problem of *nonidentification* or *underdetermination* of theories by observations (Curd and Cover, 1998). In particular, available observations are always consistent with several hypotheses that themselves are mutually inconsistent, which explains why (as Hume noted) scientific theories cannot be logically proven. In particular, consistency between a hypothesis and observations is no proof of the hypothesis, because we can always invent alternative hypotheses that are just as consistent with the observations.

In contrast, a valid observation that is inconsistent with a hypothesis implies that the hypothesis as stated is false and so refutes the hypothesis. If you wring the rooster's neck before it crows and the sun still rises, you have disproved that the rooster's crowing is a necessary cause of sunrise. Or consider a hypothetical research program to learn the boiling point of water (Magee, 1985). A scientist who boils water in an open flask and repeatedly measures the boiling point at 100°C will never, no matter how many confirmatory repetitions are involved, prove that 100°C is always the boiling point. On the other hand, merely one attempt to boil the water in a closed flask or at high altitude will refute the proposition that water always boils at 100°C.

According to Popper, science advances by a process of elimination that he called "conjecture and refutation." Scientists form hypotheses based on intuition, conjecture, and previous experience. Good scientists use deductive logic to infer predictions from the hypothesis and then compare observations with the predictions. Hypotheses whose predictions agree with observations are confirmed (Popper used the term "corroborated") only in the sense that they can continue to be used as explanations of natural phenomena. At any time, however, they may be refuted by further observations and might be replaced by other hypotheses that are more consistent with the observations. This view of scientific inference is sometimes called *refutationism* or *falsificationism.* Refutationists consider induction to be a psychologic crutch: Repeated observations did not in fact induce the formulation of a natural law, but only the belief that such a law has been found. For a refutationist, only the psychologic comfort provided by induction explains why it still has advocates.

One way to rescue the concept of induction from the stigma of pure delusion is to resurrect it as a psychologic phenomenon, as Hume and Popper claimed it was, but one that plays a legitimate role in hypothesis formation. The philosophy of conjecture and refutation places no constraints on the origin of conjectures. Even delusions are permitted as hypotheses, and therefore inductively inspired hypotheses, however psychologic, are valid starting points for scientific evaluation. This concession does not admit a logical role for induction in confirming scientific hypotheses, but it allows the process of induction to play a part, along with imagination, in the scientific cycle of conjecture and refutation.

The philosophy of conjecture and refutation has profound implications for the methodology of science. The popular concept of a scientist doggedly assembling evidence to support a favorite thesis is objectionable from the standpoint of refutationist philosophy because it encourages scientists to consider their own pet theories as their intellectual property, to be confirmed, proven, and, when all the evidence is in, cast in stone and defended as natural law. Such attitudes hinder critical evaluation, interchange, and progress. The approach of conjecture and refutation, in contrast, encourages scientists to consider multiple hypotheses and to seek crucial tests that decide between competing hypotheses by falsifying one of them. Because falsification of one or more theories is the goal, there is incentive to depersonalize the theories. Criticism leveled at a theory need not be seen as criticism of the person who proposed it. It has been suggested that the reason why certain fields of science advance rapidly while others languish is that the rapidly advancing fields are propelled by scientists

who are busy constructing and testing competing hypotheses; the other fields, in contrast, "are sick by comparison, because they have forgotten the necessity for alternative hypotheses and disproof" (Platt, 1964).

The refutationist model of science has a number of valuable lessons for research conduct, especially of the need to seek alternative explanations for observations, rather than focus on the chimera of seeking scientific "proof" for some favored theory. Nonetheless, it is vulnerable to criticisms that observations (or some would say their interpretations) are themselves laden with theory (sometimes called the *Duhem-Quine thesis*; Curd and Cover, 1998). Thus, observations can never provide the sort of definitive refutations that are the hallmark of popular accounts of refutationism. For example, there may be uncontrolled and even unimagined biases that have made our refutational observations invalid; to claim refutation is to assume as true the unprovable theory that no such bias exists. In other words, not only are theories underdetermined by observations, so are refutations, which are themselves theory-laden. The net result is that logical certainty about either the truth or falsity of an internally consistent theory is impossible (Quine, 1951).

CONSENSUS AND NATURALISM

Some 20th-century philosophers of science, most notably Thomas Kuhn (1962), emphasized the role of the scientific community in judging the validity of scientific theories. These critics of the conjecture-and-refutation model suggested that the refutation of a theory involves making a choice. Every observation is itself dependent on theories. For example, observing the moons of Jupiter through a telescope seems to us like a direct observation, but only because the theory of optics on which the telescope is based is so well accepted. When confronted with a refuting observation, a scientist faces the choice of rejecting either the validity of the theory being tested or the validity of the refuting observation, which itself must be premised on scientific theories that are not certain (Haack, 2003). Observations that are falsifying instances of theories may at times be treated as "anomalies," tolerated without falsifying the theory in the hope that the anomalies may eventually be explained. An epidemiologic example is the observation that shallow-inhaling smokers had higher lung cancer rates than deep-inhaling smokers. This anomaly was eventually explained when it was noted that lung tissue higher in the lung is more susceptible to smoking-associated lung tumors, and shallowly inhaled smoke tars tend to be deposited higher in the lung (Wald, 1985).

In other instances, anomalies may lead eventually to the overthrow of current scientific doctrine, just as Newtonian mechanics was displaced (remaining only as a first-order approximation) by relativity theory. Kuhn asserted that in every branch of science the prevailing scientific viewpoint, which he termed "normal science," occasionally undergoes major shifts that amount to scientific revolutions. These revolutions signal a decision of the scientific community to discard the scientific infrastructure rather than to falsify a new hypothesis that cannot be easily grafted onto it. Kuhn and others have argued that the consensus of the scientific community determines what is considered accepted and what is considered refuted.

Kuhn's critics characterized this description of science as one of an irrational process, "a matter for mob psychology" (Lakatos, 1970). Those who believe in a rational structure for science consider Kuhn's vision to be a regrettably real description of much of what passes for scientific activity, but not prescriptive for any good science. Although many modern philosophers reject rigid demarcations and formulations for science such as refutationism, they nonetheless maintain that science is founded on reason, albeit possibly informal common sense (Haack, 2003). Others go beyond Kuhn and maintain that attempts to impose a singular rational structure or methodology on science hobbles the imagination and is a prescription for the same sort of authoritarian repression of ideas that scientists have had to face throughout history (Feyerabend, 1975 and 1993).

The philosophic debate about Kuhn's description of science hinges on whether Kuhn meant to describe only what has happened historically in science or instead what ought to happen, an issue about which Kuhn (1970) has not been completely clear:

> Are Kuhn's [my] remarks about scientific development. . . to be read as descriptions or prescriptions? The answer, of course, is that they should be read in both ways at once. If I have a theory of how and why science works, it must necessarily have implications for the way in which scientists should behave if their enterprise is to flourish.

The idea that science is a sociologic process, whether considered descriptive or normative, is an interesting thesis, as is the idea that from observing how scientists work we can learn about how scientists ought to work. The latter idea has led to the development of *naturalistic* philosophy of science, or "science studies," which examines scientific developments for clues about what sort of methods scientists need and develop for successful discovery and invention (Callebaut, 1993; Giere, 1999).

Regardless of philosophical developments, we suspect that most epidemiologists (and most scientists) will continue to function as if the following classical view is correct: The ultimate goal of scientific inference is to capture some objective truths about the material world in which we live, and any theory of inference should ideally be evaluated by how well it leads us to these truths. This ideal is impossible to operationalize, however, for if we ever find any ultimate truths, we will have no way of knowing that for certain. Thus, those holding the view that scientific truth is not arbitrary nevertheless concede that our knowledge of these truths will always be tentative. For refutationists, this tentativeness has an asymmetric quality, but that asymmetry is less marked for others. We may believe that we know a theory is false because it consistently fails the tests we put it through, but our tests could be faulty, given that they involve imperfect reasoning and sense perception. Neither can we know that a theory is true, even if it passes every test we can devise, for it may fail a test that is as yet undevised.

Few, if any, would disagree that a theory of inference should be evaluated at least in part by how well it leads us to detect errors in our hypotheses and observations. There are, however, many other inferential activities besides evaluation of hypotheses, such as prediction or forecasting of events, and subsequent attempts to control events (which of course requires causal information). Statisticians rather than philosophers have more often confronted these problems in practice, so it should not be surprising that the major philosophies concerned with these problems emerged from statistics rather than philosophy.

BAYESIANISM

There is another philosophy of inference that, like most, holds an objective view of scientific truth and a view of knowledge as tentative or uncertain, but that focuses on evaluation of knowledge rather than truth. Like refutationism, the modern form of this philosophy evolved from the writings of 18th-century thinkers. The focal arguments first appeared in a pivotal essay by the Reverend Thomas Bayes (1764), and hence the philosophy is usually referred to as Bayesianism (Howson and Urbach, 1993), and it was the renowned French mathematician and scientist Pierre Simon de Laplace who first gave it an applied statistical format. Nonetheless, it did not reach a complete expression until after World War I, most notably in the writings of Ramsey (1931) and DeFinetti (1937); and, like refutationism, it did not begin to appear in epidemiology until the 1970s (e.g., Cornfield, 1976).

The central problem addressed by Bayesianism is the following: In classical logic, a deductive argument can provide no information about the truth or falsity of a scientific hypothesis unless you can be 100% certain about the truth of the premises of the argument. Consider the logical argument called *modus tollens:* "If H implies B, and B is false, then H must be false." This argument is logically valid, but the conclusion follows only on the assumptions that the premises "H implies B" and "B is false" are true statements. If these premises are statements about the physical world, we cannot possibly know them to be correct with 100% certainty, because all observations are subject to error. Furthermore, the claim that "H implies B" will often depend on its own chain of deductions, each with its own premises of which we cannot be certain.

For example, if H is "Television viewing causes homicides" and B is "Homicide rates are highest where televisions are most common," the first premise used in *modus tollens* to test the hypothesis that television viewing causes homicides will be: "If television viewing causes homicides, homicide rates are highest where televisions are most common." The validity of this premise is doubtful—after all, even if television does cause homicides, homicide rates may be low where televisions are common because of socioeconomic advantages in those areas.

Continuing to reason in this fashion, we could arrive at a more pessimistic state than even Hume imagined. Not only is induction without logical foundation, *deduction* has limited scientific utility because we cannot ensure the truth of all the premises, even if a logical argument is valid.

The Bayesian answer to this problem is partial in that it makes a severe demand on the scientist and puts a severe limitation on the results. It says roughly this: If you can assign a degree of certainty, or personal probability, to the premises of your valid argument, you may use any and all the rules of probability theory to derive a certainty for the conclusion, and this certainty will be a logically valid consequence of your original certainties. An inescapable fact is that your concluding certainty, or *posterior probability*, may depend heavily on what you used as initial certainties, or *prior probabilities*. If those initial certainties are not the same as those of a colleague, that colleague may very well assign a certainty to the conclusion different from the one you derived. With the accumulation of consistent evidence, however, the data can usually force even extremely disparate priors to converge into similar posterior probabilities.

Because the posterior probabilities emanating from a Bayesian inference depend on the person supplying the initial certainties and so may vary across individuals, the inferences are said to be subjective. This subjectivity of Bayesian inference is often mistaken for a subjective treatment of truth. Not only is such a view of Bayesianism incorrect, it is diametrically opposed to Bayesian philosophy. The Bayesian approach represents a constructive attempt to deal with the dilemma that scientific laws and facts should not be treated as known with certainty, whereas classic deductive logic yields conclusions only when some law, fact, or connection is asserted with 100% certainty.

A common criticism of Bayesian philosophy is that it diverts attention away from the classic goals of science, such as the discovery of how the world works, toward psychologic states of mind called "certainties," "subjective probabilities," or "degrees of belief" (Popper, 1959). This criticism, however, fails to recognize the importance of a scientist's state of mind in determining what theories to test and what tests to apply, the consequent influence of those states on the store of data available for inference, and the influence of the data on the states of mind.

Another reply to this criticism is that scientists already use data to influence their degrees of belief, and they are not shy about expressing those degrees of certainty. The problem is that the conventional process is informal, intuitive, and ineffable, and therefore not subject to critical scrutiny; at its worst, it often amounts to nothing more than the experts announcing that they have seen the evidence and here is how certain they are. How they reached this certainty is left unclear, or, put another way, is not "transparent." The problem is that no one, even an expert, is very good at informally and intuitively formulating certainties that predict facts and future events well (Kahneman et al., 1982; Gilovich, 1993; Piattelli-Palmarini, 1994; Gilovich et al., 2002). One reason for this problem is that biases and prior prejudices can easily creep into expert judgments. Bayesian methods force experts to "put their cards on the table" and specify explicitly the strength of their prior beliefs and why they have such beliefs, defend those specifications against arguments and evidence, and update their degrees of certainty with new evidence in ways that do not violate probability logic.

In any research context, there will be an unlimited number of hypotheses that could explain an observed phenomenon. Some argue that progress is best aided by severely testing (empirically challenging) those explanations that seem most probable in light of past research, so that shortcomings of currently "received" theories can be most rapidly discovered. Indeed, much research in certain fields takes this form, as when theoretical predictions of particle mass are put to ever more precise tests in physics experiments. This process does not involve mere improved repetition of past studies. Rather, it involves tests of previously untested but important predictions of the theory. Moreover, there is an imperative to make the basis for prior beliefs criticizable and defensible. That prior probabilities can differ among persons does not mean that all such beliefs are based on the same information, nor that all are equally tenable.

Probabilities of auxiliary hypotheses are also important in study design and interpretation. Failure of a theory to pass a test can lead to rejection of the theory more rapidly when the auxiliary hypotheses on which the test depends possess high probability. This observation provides a rationale for preferring "nested" case-control studies (in which controls are selected from a roster of the source population for the cases) to "hospital-based" case-control studies (in which the controls are "selected" by the occurrence or diagnosis of one or more diseases other than the case-defining disease), because the former have fewer mechanisms for biased subject selection and hence are given a higher probability of unbiased subject selection.

Even if one disputes the above arguments, most epidemiologists desire some way of expressing the varying degrees of certainty about possible values of an effect measure in light of available data. Such expressions must inevitably be derived in the face of considerable uncertainty about

methodologic details and various events that led to the available data and can be extremely sensitive to the reasoning used in its derivation. For example, as we shall discuss at greater length in Chapter 19, conventional confidence intervals quantify only random error under often questionable assumptions and so should not be interpreted as measures of total uncertainty, particularly for nonexperimental studies. As noted earlier, most people, including scientists, reason poorly in the face of uncertainty. At the very least, subjective Bayesian philosophy provides a methodology for sound reasoning under uncertainty and, in particular, provides many warnings against being overly certain about one's conclusions (Greenland 1998a, 1988b, 2006a; see also Chapters 18 and 19).

Such warnings are echoed in refutationist philosophy. As Peter Medawar (1979) put it, "I cannot give any scientist of any age better advice than this: the intensity of the conviction that a hypothesis is true has no bearing on whether it is true or not." We would add two points. First, the intensity of conviction that a hypothesis is false has no bearing on whether it is false or not. Second, Bayesian methods do not mistake beliefs for evidence. They use evidence to modify beliefs, which scientists routinely do in any event, but often in implicit, intuitive, and incoherent ways.

IMPOSSIBILITY OF SCIENTIFIC PROOF

Vigorous debate is a characteristic of modern scientific philosophy, no less in epidemiology than in other areas (Rothman, 1988). Can divergent philosophies of science be reconciled? Haack (2003) suggested that the scientific enterprise is akin to solving a vast, collective crossword puzzle. In areas in which the evidence is tightly interlocking, there is more reason to place confidence in the answers, but in areas with scant information, the theories may be little better than informed guesses. Of the scientific method, Haack (2003) said that "there is less to the 'scientific method' than meets the eye. Is scientific inquiry categorically different from other kinds? No. Scientific inquiry is continuous with everyday empirical inquiry—only more so."

Perhaps the most important common thread that emerges from the debated philosophies is that proof is impossible in empirical science. This simple fact is especially important to observational epidemiologists, who often face the criticism that proof is impossible in epidemiology, with the implication that it is possible in other scientific disciplines. Such criticism may stem from a view that experiments are the definitive source of scientific knowledge. That view is mistaken on at least two counts. First, the nonexperimental nature of a science does not preclude impressive scientific discoveries; the myriad examples include plate tectonics, the evolution of species, planets orbiting other stars, and the effects of cigarette smoking on human health. Even when they are possible, experiments (including randomized trials) do not provide anything approaching proof and in fact may be controversial, contradictory, or nonreproducible. If randomized clinical trials provided proof, we would never need to do more than one of them on a given hypothesis. Neither physical nor experimental science is immune to such problems, as demonstrated by episodes such as the experimental "discovery" (later refuted) of cold fusion (Taubes, 1993).

Some experimental scientists hold that epidemiologic relations are only suggestive and believe that detailed laboratory study of mechanisms within single individuals can reveal cause–effect relations with certainty. This view overlooks the fact that *all* relations are suggestive in exactly the manner discussed by Hume. Even the most careful and detailed mechanistic dissection of individual events cannot provide more than associations, albeit at a finer level. Laboratory studies often involve a degree of observer control that cannot be approached in epidemiology; it is only this control, not the level of observation, that can strengthen the inferences from laboratory studies. And again, such control is no guarantee against error. In addition, neither scientists nor decision makers are often highly persuaded when only mechanistic evidence from the laboratory is available.

All of the fruits of scientific work, in epidemiology or other disciplines, are at best only tentative formulations of a description of nature, even when the work itself is carried out without mistakes. The tentativeness of our knowledge does not prevent practical applications, but it should keep us skeptical and critical, not only of everyone else's work, but of our own as well. Sometimes etiologic hypotheses enjoy an extremely high, universally or almost universally shared, degree of certainty. The hypothesis that cigarette smoking causes lung cancer is one of the best-known examples. These hypotheses rise above "tentative" acceptance and are the closest we can come to "proof." But even

these hypotheses are not "proved" with the degree of absolute certainty that accompanies the proof of a mathematical theorem.

CAUSAL INFERENCE IN EPIDEMIOLOGY

Etiologic knowledge about epidemiologic hypotheses is often scant, making the hypotheses themselves at times little more than vague statements of causal association between exposure and disease, such as "smoking causes cardiovascular disease." These vague hypotheses have only vague consequences that can be difficult to test. To cope with this vagueness, epidemiologists usually focus on testing the negation of the causal hypothesis, that is, the null hypothesis that the exposure does *not* have a causal relation to disease. Then, any observed association can potentially refute the hypothesis, subject to the assumption (auxiliary hypothesis) that biases and chance fluctuations are not solely responsible for the observation.

TESTS OF COMPETING EPIDEMIOLOGIC THEORIES

If the causal mechanism is stated specifically enough, epidemiologic observations can provide crucial tests of competing, non-null causal hypotheses. For example, when toxic-shock syndrome was first studied, there were two competing hypotheses about the causal agent. Under one hypothesis, it was a chemical in the tampon, so that women using tampons were exposed to the agent directly from the tampon. Under the other hypothesis, the tampon acted as a culture medium for staphylococci that produced a toxin. Both hypotheses explained the relation of toxic-shock occurrence to tampon use. The two hypotheses, however, led to opposite predictions about the relation between the frequency of changing tampons and the rate of toxic shock. Under the hypothesis of a chemical agent, more frequent changing of the tampon would lead to more exposure to the agent and possible absorption of a greater overall dose. This hypothesis predicted that women who changed tampons more frequently would have a higher rate than women who changed tampons infrequently. The culture-medium hypothesis predicts that women who change tampons frequently would have a lower rate than those who change tampons less frequently, because a short duration of use for each tampon would prevent the staphylococci from multiplying enough to produce a damaging dose of toxin. Thus, epidemiologic research, by showing that infrequent changing of tampons was associated with a higher rate of toxic shock, refuted the chemical theory in the form presented. There was, however, a third hypothesis that a chemical in some tampons (e.g., oxygen content) improved their performance as culture media. This chemical-promotor hypothesis made the same prediction about the association with frequency of changing tampons as the microbial toxin hypothesis (Lanes and Rothman, 1990).

Another example of a theory that can be easily tested by epidemiologic data relates to the observation that women who took replacement estrogen therapy had a considerably elevated rate of endometrial cancer. Horwitz and Feinstein (1978) conjectured a competing theory to explain the association: They proposed that women taking estrogen experienced symptoms such as bleeding that induced them to consult a physician. The resulting diagnostic workup led to the detection of endometrial cancer at an earlier stage in these women, as compared with women who were not taking estrogens. Horwitz and Feinstein argued that the association arose from this detection bias, claiming that without the bleeding-induced workup, many of these cancers would not have been detected at all. Many epidemiologic observations were used to evaluate these competing hypotheses. The detection-bias theory predicted that women who had used estrogens for only a short time would have the greatest elevation in their rate, as the symptoms related to estrogen use that led to the medical consultation tended to appear soon after use began. Because the association of recent estrogen use and endometrial cancer was the same in both long- and short-term estrogen users, the detection-bias theory was refuted as an explanation for all but a small fraction of endometrial cancer cases occurring after estrogen use. Refutation of the detection-bias theory also depended on many other observations. Especially important was the theory's implication that there must be a huge reservoir of undetected endometrial cancer in the typical population of women to account for the much greater rate observed in estrogen users, an implication that was not borne out by further observations (Hutchison and Rothman, 1978).

The endometrial cancer example illustrates a critical point in understanding the process of causal inference in epidemiologic studies: Many of the hypotheses being evaluated in the interpretation of epidemiologic studies are auxiliary hypotheses in the sense that they are independent of the presence, absence, or direction of any causal connection between the study exposure and the disease. For example, explanations of how specific types of bias could have distorted an association between exposure and disease are the usual alternatives to the primary study hypothesis. Much of the interpretation of epidemiologic studies amounts to the testing of such auxiliary explanations for observed associations.

CAUSAL CRITERIA

In practice, how do epidemiologists separate causal from noncausal explanations? Despite philosophic criticisms of inductive inference, inductively oriented considerations are often used as criteria for making such inferences (Weed and Gorelic, 1996). If a set of necessary and sufficient causal criteria could be used to distinguish causal from noncausal relations in epidemiologic studies, the job of the scientist would be eased considerably. With such criteria, all the concerns about the logic or lack thereof in causal inference could be subsumed: It would only be necessary to consult the checklist of criteria to see if a relation were causal. We know from the philosophy reviewed earlier that a set of sufficient criteria does not exist. Nevertheless, lists of causal criteria have become popular, possibly because they seem to provide a road map through complicated territory, and perhaps because they suggest hypotheses to be evaluated in a given problem.

A commonly used set of criteria was based on a list of considerations or "viewpoints" proposed by Sir Austin Bradford Hill (1965). Hill's list was an expansion of a list offered previously in the landmark U.S. Surgeon General's report *Smoking and Health* (1964), which in turn was anticipated by the inductive canons of John Stuart Mill (1862) and the rules given by Hume (1739). Subsequently, others, especially Susser, have further developed causal considerations (Kaufman and Poole, 2000).

Hill suggested that the following considerations in attempting to distinguish causal from noncausal associations that were already "perfectly clear-cut and beyond what we would care to attribute to the play of chance": (1) strength, (2) consistency, (3) specificity, (4) temporality, (5) biologic gradient, (6) plausibility, (7) coherence, (8) experimental evidence, and (9) analogy. Hill emphasized that causal inferences cannot be based on a set of rules, condemned emphasis on statistical significance testing, and recognized the importance of many other factors in decision making (Phillips and Goodman, 2004). Nonetheless, the misguided but popular view that his considerations should be used as criteria for causal inference makes it necessary to examine them in detail.

Strength

Hill argued that strong associations are particularly compelling because, for weaker associations, it is "easier" to imagine what today we would call an unmeasured confounder that might be responsible for the association. Several years earlier, Cornfield et al. (1959) drew similar conclusions. They concentrated on a single hypothetical confounder that, by itself, would explain entirely an observed association. They expressed a strong preference for ratio measures of strength, as opposed to difference measures, and focused on how the observed estimate of a risk ratio provides a minimum for the association that a completely explanatory confounder must have with the exposure (rather than a minimum for the confounder–disease association). Of special importance, Cornfield et al. acknowledged that having only a weak association does not rule out a causal connection (Rothman and Poole, 1988). Today, some associations, such as those between smoking and cardiovascular disease or between environmental tobacco smoke and lung cancer, are accepted by most as causal even though the associations are considered weak.

Counterexamples of strong but noncausal associations are also not hard to find; any study with strong confounding illustrates the phenomenon. For example, consider the strong relation between Down syndrome and birth rank, which is confounded by the relation between Down syndrome and maternal age. Of course, once the confounding factor is identified, the association is diminished by controlling for the factor.

These examples remind us that a strong association is neither necessary nor sufficient for causality, and that weakness is neither necessary nor sufficient for absence of causality. A strong association

bears only on hypotheses that the association is entirely or partially due to unmeasured confounders or other source of modest bias.

Consistency

To most observers, consistency refers to the repeated observation of an association in different populations under different circumstances. Lack of consistency, however, does not rule out a causal association, because some effects are produced by their causes only under unusual circumstances. More precisely, the effect of a causal agent cannot occur unless the complementary component causes act or have already acted to complete a sufficient cause. These conditions will not always be met. Thus, transfusions can cause infection with the human immunodeficiency virus, but they do not always do so: The virus must also be present. Tampon use can cause toxic-shock syndrome, but only rarely, when certain other, perhaps unknown, conditions are met. Consistency is apparent only after all the relevant details of a causal mechanism are understood, which is to say very seldom. Furthermore, even studies of exactly the same phenomena can be expected to yield different results simply because they differ in their methods and random errors. Consistency serves only to rule out hypotheses that the association is attributable to some factor that varies across studies.

One mistake in implementing the consistency criterion is so common that it deserves special mention. It is sometimes claimed that a literature or set of results is inconsistent simply because some results are "statistically significant" and some are not. This sort of evaluation is completely fallacious even if one accepts the use of significance testing methods. The results (effect estimates) from a set of studies could all be identical even if many were significant and many were not, the difference in significance arising solely because of differences in the standard errors or sizes of the studies. Conversely, the results could be significantly in conflict even if all were all were nonsignificant individually, simply because in aggregate an effect could be apparent in some subgroups but not others (see Chapter 33). The fallacy of judging consistency by comparing *P*-values or statistical significance is not eliminated by "standardizing" estimates (i.e., dividing them by the standard deviation of the outcome, multiplying them by the standard deviation of the exposure, or both); in fact it is worsened, as such standardization can create differences where none exists, or mask true differences (Greenland et al., 1986, 1991; see Chapters 21 and 33).

Specificity

The criterion of specificity has two variants. One is that a cause leads to a single effect, not multiple effects. The other is that an effect has one cause, not multiple causes. Hill mentioned both of them. The former criterion, specificity of effects, was used as an argument in favor of a causal interpretation of the association between smoking and lung cancer and, in an act of circular reasoning, in favor of ratio comparisons and not differences as the appropriate measures of strength. When ratio measures were examined, the association of smoking to diseases looked "quantitatively specific" to lung cancer. When difference measures were examined, the association appeared to be nonspecific, with several diseases (other cancers, coronary heart disease, etc.) being at least as strongly associated with smoking as lung cancer was. Today we know that smoking affects the risk of many diseases and that the difference comparisons were accurately portraying this lack of specificity. Unfortunately, however, the historical episode of the debate over smoking and health is often cited today as justification for the specificity criterion and for using ratio comparisons to measure strength of association. The proper lessons to learn from that episode should be just the opposite.

Weiss (2002) argued that specificity can be used to distinguish some causal hypotheses from noncausal hypotheses, when the causal hypothesis predicts a relation with one outcome but no relation with another outcome. His argument is persuasive when, in addition to the causal hypothesis, one has an alternative noncausal hypothesis that predicts a nonspecific association. Weiss offered the example of screening sigmoidoscopy, which was associated in case-control studies with a 50% to 70% reduction in mortality from distal tumors of the rectum and tumors of the distal colon, within the reach of the sigmoidoscope, but no reduction in mortality from tumors elsewhere in the colon. If the effect of screening sigmoidoscopy were not specific to the distal colon tumors, it would lend support not to all noncausal theories to explain the association, as Weiss suggested, but only to those noncausal theories that would have predicted a nonspecific association. Thus, specificity can

come into play when it can be logically deduced from the causal hypothesis in question and when nonspecificity can be logically deduced from one or more noncausal hypotheses.

Temporality

Temporality refers to the necessity that the cause precede the effect in time. This criterion is inarguable, insofar as any claimed observation of causation must involve the putative cause C preceding the putative effect D. It does *not,* however, follow that a reverse time order is evidence against the hypothesis that C can cause D. Rather, observations in which C followed D merely show that C could not have caused D in these instances; they provide no evidence for or against the hypothesis that C can cause D in those instances in which it precedes D. Only if it is found that C cannot precede D can we dispense with the causal hypothesis that C *could* cause D.

Biologic Gradient

Biologic gradient refers to the presence of a dose–response or exposure–response curve with an expected shape. Although Hill referred to a "linear" gradient, without specifying the scale, a linear gradient on one scale, such as the risk, can be distinctly nonlinear on another scale, such as the log risk, the odds, or the log odds. We might relax the expectation from linear to strictly monotonic (steadily increasing or decreasing) or even further merely to monotonic (a gradient that never changes direction). For example, more smoking means more carcinogen exposure and more tissue damage, hence more opportunity for carcinogenesis. Some causal associations, however, show a rapid increase in response (an approximate threshold effect) rather than a strictly monotonic trend. An example is the association between DES and adenocarcinoma of the vagina. A possible explanation is that the doses of DES that were administered were all sufficiently great to produce the maximum effect from DES. Under this hypothesis, for all those exposed to DES, the development of disease would depend entirely on other component causes.

The somewhat controversial topic of alcohol consumption and mortality is another example. Death rates are higher among nondrinkers than among moderate drinkers, but they ascend to the highest levels for heavy drinkers. There is considerable debate about which parts of the J-shaped dose–response curve are causally related to alcohol consumption and which parts are noncausal artifacts stemming from confounding or other biases. Some studies appear to find only an increasing relation between alcohol consumption and mortality, possibly because the categories of alcohol consumption are too broad to distinguish different rates among moderate drinkers and nondrinkers, or possibly because they have less confounding at the lower end of the consumption scale.

Associations that do show a monotonic trend in disease frequency with increasing levels of exposure are not necessarily causal. Confounding can result in a monotonic relation between a noncausal risk factor and disease if the confounding factor itself demonstrates a biologic gradient in its relation with disease. The relation between birth rank and Down syndrome mentioned earlier shows a strong biologic gradient that merely reflects the progressive relation between maternal age and occurrence of Down syndrome.

These issues imply that the existence of a monotonic association is neither necessary nor sufficient for a causal relation. A nonmonotonic relation only refutes those causal hypotheses specific enough to predict a monotonic dose–response curve.

Plausibility

Plausibility refers to the scientific plausibility of an association. More than any other criterion, this one shows how narrowly systems of causal criteria are focused on epidemiology. The starting point is an epidemiologic association. In asking whether it is causal or not, one of the considerations we take into account is its plausibility. From a less parochial perspective, the entire enterprise of causal inference would be viewed as the act of determining how plausible a causal *hypothesis* is. One of the considerations we would take into account would be epidemiologic associations, if they are available. Often they are not, but causal inference must be done nevertheless, with inputs from toxicology, pharmacology, basic biology, and other sciences.

Just as epidemiology is not essential for causal inference, plausibility can change with the times. Sartwell (1960) emphasized this point, citing remarks of Cheever in 1861, who had been commenting on the etiology of typhus before its mode of transmission (via body lice) was known:

> It could be no more ridiculous for the stranger who passed the night in the steerage of an emigrant ship to ascribe the typhus, which he there contracted, to the vermin with which bodies of the sick might be infested. An adequate cause, one reasonable in itself, must correct the coincidences of simple experience.

What was to Cheever an implausible explanation turned out to be the correct explanation, because it was indeed the vermin that caused the typhus infection. Such is the problem with plausibility: It is too often based not on logic or data, but only on prior beliefs. This is not to say that biologic knowledge should be discounted when a new hypothesis is being evaluated, but only to point out the difficulty in applying that knowledge.

The Bayesian approach to inference attempts to deal with this problem by requiring that one quantify, on a probability (0 to 1) scale, the certainty that one has in prior beliefs, as well as in new hypotheses. This quantification displays the dogmatism or open-mindedness of the analyst in a public fashion, with certainty values near 1 or 0 betraying a strong commitment of the analyst for or against a hypothesis. It can also provide a means of testing those quantified beliefs against new evidence (Howson and Urbach, 1993). Nevertheless, no approach can transform plausibility into an objective causal criterion.

Coherence

Taken from the U.S. Surgeon General's *Smoking and Health* (1964), the term *coherence* implies that a cause-and-effect interpretation for an association does not conflict with what is known of the natural history and biology of the disease. The examples Hill gave for coherence, such as the histopathologic effect of smoking on bronchial epithelium (in reference to the association between smoking and lung cancer) or the difference in lung cancer incidence by sex, could reasonably be considered examples of plausibility, as well as coherence; the distinction appears to be a fine one. Hill emphasized that the absence of coherent information, as distinguished, apparently, from the presence of conflicting information, should not be taken as evidence against an association being considered causal. On the other hand, the presence of conflicting information may indeed refute a hypothesis, but one must always remember that the conflicting information may be mistaken or misinterpreted. An example mentioned earlier is the "inhalation anomaly" in smoking and lung cancer, the fact that the excess of lung cancers seen among smokers seemed to be concentrated at sites in the upper airways of the lung. Several observers interpreted this anomaly as evidence that cigarettes were not responsible for the excess. Other observations, however, suggested that cigarette-borne carcinogens were deposited preferentially where the excess was observed, and so the anomaly was in fact consistent with a causal role for cigarettes (Wald, 1985).

Experimental Evidence

To different observers, experimental evidence can refer to clinical trials, to laboratory experiments with rodents or other nonhuman organisms, or to both. Evidence from human experiments, however, is seldom available for epidemiologic research questions, and animal evidence relates to different species and usually to levels of exposure very different from those that humans experience. Uncertainty in extrapolations from animals to humans often dominates the uncertainty of quantitative risk assessments (Freedman and Zeisel, 1988; Crouch et al., 1997).

To Hill, however, experimental evidence meant something else: the "experimental, or semi-experimental evidence" obtained from reducing or eliminating a putatively harmful exposure and seeing if the frequency of disease subsequently declines. He called this the strongest possible evidence of causality that can be obtained. It can be faulty, however, as the "semi-experimental" approach is nothing more than a "before-and-after" time trend analysis, which can be confounded or otherwise biased by a host of concomitant secular changes. Moreover, even if the removal of exposure does causally reduce the frequency of disease, it might not be for the etiologic reason hypothesized. The draining of a swamp near a city, for instance, would predictably and causally reduce the rate of yellow fever or malaria in that city the following summer. But it would be a mistake to call this observation the strongest possible evidence of a causal role of miasmas (Poole, 1999).

Analogy

Whatever insight might be derived from analogy is handicapped by the inventive imagination of scientists who can find analogies everywhere. At best, analogy provides a source of more elaborate hypotheses about the associations under study; absence of such analogies reflects only lack of imagination or experience, not falsity of the hypothesis.

We might find naive Hill's examples in which reasoning by analogy from the thalidomide and rubella tragedies made it more likely to him that other medicines and infections might cause other birth defects. But such reasoning is common; we suspect most people find it more credible that smoking might cause, say, stomach cancer, because of its associations, some widely accepted as causal, with cancers in other internal and gastrointestinal organs. Here we see how the analogy criterion can be at odds with either of the two specificity criteria. The more apt the analogy, the less specific are the effects of a cause or the less specific the causes of an effect.

Summary

As is evident, the standards of epidemiologic evidence offered by Hill are saddled with reservations and exceptions. Hill himself was ambivalent about their utility. He did not use the word *criteria* in the speech. He called them "viewpoints" or "perspectives." On the one hand, he asked, "In what circumstances can we pass from this observed *association* to a verdict of *causation*?" (emphasis in original). Yet, despite speaking of verdicts on causation, he disagreed that any "hard-and-fast rules of evidence" existed by which to judge causation: "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non" (Hill, 1965).

Actually, as noted above, the fourth viewpoint, temporality, is *a sine qua non* for causal explanations of observed associations. Nonetheless, it does not bear on the hypothesis that an exposure is capable of causing a disease in situations as yet unobserved (whether in the past or the future). For suppose every exposed case of disease ever reported had received the exposure after developing the disease. This reversed temporal relation would imply that exposure had not caused disease among these reported cases, and thus would refute the hypothesis that it had. Nonetheless, it would *not* refute the hypothesis that the exposure is *capable* of causing the disease, or that it had caused the disease in unobserved cases. It would mean only that we have no worthwhile epidemiologic evidence relevant to that hypothesis, for we had not yet seen what became of those exposed before disease occurred relative to those unexposed. Furthermore, what appears to be a causal sequence could represent reverse causation if preclinical symptoms of the disease lead to exposure, and then overt disease follows, as when patients in pain take analgesics, which may be the result of disease that is later diagnosed, rather than a cause.

Other than temporality, there is no necessary or sufficient criterion for determining whether an observed association is causal. Only when a causal hypothesis is elaborated to the extent that one can predict from it a particular form of consistency, specificity, biologic gradient, and so forth, can "causal criteria" come into play in evaluating causal hypotheses, and even then they do not come into play in evaluating the general hypothesis per se, but only some specific causal hypotheses, leaving others untested.

This conclusion accords with the views of Hume and many others that causal inferences cannot attain the certainty of logical deductions. Although some scientists continue to develop causal considerations as aids to inference (Susser, 1991), others argue that it is detrimental to cloud the inferential process by considering checklist criteria (Lanes and Poole, 1984). An intermediate, refutationist approach seeks to transform proposed criteria into deductive tests of causal hypotheses (Maclure, 1985; Weed, 1986). Such an approach helps avoid the temptation to use causal criteria simply to buttress pet theories at hand, and instead allows epidemiologists to focus on evaluating competing causal theories using crucial observations. Although this refutationist approach to causal inference may seem at odds with the common implementation of Hill's viewpoints, it actually seeks to answer the fundamental question posed by Hill, and the ultimate purpose of the viewpoints he promulgated:

> What [the nine viewpoints] can do, with greater or less strength, is to help us to make up our minds on the fundamental question—is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect? (Hill, 1965)

The crucial phrase "equally or more likely than cause and effect" suggests to us a subjective assessment of the certainty, or probability of the causal hypothesis at issue relative to another hypothesis. Although Hill wrote at a time when expressing uncertainty as a probability was unpopular in statistics, it appears from his statement that, for him, causal inference is a subjective matter of degree of personal belief, certainty, or conviction. In any event, this view is precisely that of subjective Bayesian statistics (Chapter 18).

It is unsurprising that case studies (e.g., Weed and Gorelick, 1996) and surveys of epidemiologists (Holman et al., 2001) show, contrary to the rhetoric that often attends invocations of causal criteria, that epidemiologists have *not* agreed on a set of causal criteria or on how to apply them. In one study in which epidemiologists were asked to employ causal criteria to fictional summaries of epidemiologic literatures, the agreement was only slightly greater than would have been expected by chance (Holman et al., 2001). The typical use of causal criteria is to make a case for a position for or against causality that has been arrived at by other, unstated means. Authors pick and choose among the criteria they deploy, and define and weight them in *ad hoc* ways that depend only on the exigencies of the discussion at hand. In this sense, causal criteria appear to function less like standards or principles and more like values (Poole, 2001b), which vary across individual scientists and even vary within the work of a single scientist, depending on the context and time. Thus universal and objective causal criteria, if they exist, have yet to be identified.

CHAPTER 8

Case-Control Studies

Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash

The use and understanding of case-control studies is one of the most important methodologic developments of modern epidemiology. Conceptually, there are clear links from randomized experiments to nonrandomized cohort studies, and from nonrandomized cohort studies to case-control studies. Case-control studies nevertheless differ enough from the scientific paradigm of experimentation that a casual approach to their conduct and interpretation invites misconception. In this chapter we review case-control study designs and contrast their advantages and disadvantages with cohort designs. We also consider variants of the basic case-control study design.

Conventional wisdom about case-control studies is that they do not yield estimates of effect that are as valid as measures obtained from cohort studies. This thinking may reflect common misunderstandings in conceptualizing case-control studies, which will be clarified later. It may also reflect concern about biased exposure information and selection in case-control studies. For example, if exposure information comes from interviews, cases will usually have reported the exposure information after learning of their diagnosis. Diagnosis may affect reporting in a number of ways, for example, by improving memory, thus enhancing sensitivity among cases, or by provoking more false memory of exposure, thus reducing specificity among cases. Furthermore, the disease may itself cloud memory and thus reduce sensitivity. These phenomena are examples of *recall bias*. Disease cannot affect exposure information collected before the disease occurred, however. Thus exposure information taken from records created before the disease occurs will not be subject to recall bias, regardless of whether the study is a cohort or a case-control design.

Conversely, cohort studies are not immune from problems often thought to be particular to case-control studies. For example, while a cohort study may gather information on exposure for an entire source population at the outset of the study, it still requires tracing of subjects to ascertain

exposure variation and outcomes. If the success of this tracing is related to the exposure and the outcome, the resulting selection bias will behave analogously to that often raised as a concern in case-control studies (Greenland, 1977; Chapter 12). Similarly, cohort studies sometimes use recall to reconstruct or impute exposure history (retrospective evaluation) and are vulnerable to recall bias if this reconstruction is done after disease occurrence. Thus, although more opportunity for recall and selection bias may arise in typical case-control studies than in typical prospective cohort studies, each study must be considered in detail to evaluate its vulnerability to bias, regardless of its design.

Conventional wisdom also holds that cohort studies are useful for evaluating the range of effects related to a single exposure, whereas case-control studies provide information only about the one disease that afflicts the cases. This thinking conflicts with the idea that case-control studies can be viewed simply as more efficient cohort studies. Just as one can choose to measure more than one disease outcome in a cohort study, it is possible to conduct a set of case-control studies nested within the same population using several disease outcomes as the case series. The case-cohort study (see below) is particularly well suited to this task, allowing one control group to be compared with several series of cases. Whether or not the case-cohort design is the form of case-control study that is used, case-control studies do not have to be characterized as being limited with respect to the number of disease outcomes that can be studied.

For diseases that are sufficiently rare, cohort studies become impractical and case-control studies become the only useful alternative. On the other hand, if exposure is rare, ordinary case-control studies are inefficient, and one must use methods that selectively recruit additional exposed subjects, such as special cohort studies or two-stage designs. If both the exposure and the outcome are rare, two-stage designs may be the only informative option, as they employ oversampling of both exposed and diseased subjects.

As understanding of the principles of case-control studies has progressed, the reputation of case-control studies has also improved. Formerly, it was common to hear case-control studies referred to disparagingly as "retrospective" studies, a term that should apply to only some case-control studies and applies as well to some cohort studies (see Chapter 6). Although case-control studies do present more opportunities for bias and mistaken inference than cohort studies, these opportunities come as a result of the relative ease with which a case-control study can be mounted. Because it need not be extremely expensive or time-consuming to conduct a case-control study, many studies have been conducted by naive investigators who do not understand or implement the basic principles of valid case-control design. Occasionally, such haphazard research can produce valuable results, but often the results are wrong because basic principles have been violated. The bad reputation once suffered by case-control studies stems more from instances of poor conduct and overinterpretation of results than from any inherent weakness in the approach.

Ideally, a case-control study can be conceptualized as a more efficient version of a corresponding cohort study. Under this conceptualization, the cases in the case-control study are the same cases as would ordinarily be included in the cohort study. Rather than including all of the experience of the source population that gave rise to the cases (the study base), as would be the usual practice in a cohort design, controls are selected from the source population. Wacholder (1996) describes this paradigm of the case-control study as a cohort study with data missing at random and by design. The sampling of controls from the population that gave rise to the cases affords the efficiency gain of a case-control design over a cohort design. The controls provide an estimate of the prevalence of the exposure and covariates in the source population. When controls are selected from members of the population who were at risk for disease at the beginning of the study's follow-up period, the case-control odds ratio estimates the risk ratio that would be obtained from a cohort design. When controls are selected from members of the population who were noncases at the times that each case occurs, or otherwise in proportion to the person-time accumulated by the cohort, the case-control odds ratio estimates the rate ratio that would be obtained from a cohort design. Finally, when controls are selected from members of the population who were noncases at the end of the study's follow-up period, the case-control odds ratio estimates the incidence odds ratio that would be obtained from a cohort design. With each control-selection strategy, the odds-ratio calculation is the same, but the measure of effect estimated by the odds ratio differs. Study designs that implement each of these control selection paradigms will be discussed after topics that are common to all designs.

COMMON ELEMENTS OF CASE-CONTROL STUDIES

In a cohort study, the numerator and denominator of each disease frequency (incidence proportion, incidence rate, or incidence odds) are measured, which requires enumerating the entire population and keeping it under surveillance—or using an existing registry—to identify cases over the follow-up period. A valid case-control study observes the population more efficiently by using a control series in place of complete assessment of the denominators of the disease frequencies. The cases in a case-control study should be the same people who would be considered cases in a cohort study of the same population.

PSEUDO-FREQUENCIES AND THE ODDS RATIO

The primary goal for control selection is that the exposure distribution among controls be the same as it is in the source population of cases. The rationale for this goal is that, if it is met, we can use the control series in place of the denominator information in measures of disease frequency to determine the ratio of the disease frequency in exposed people relative to that among unexposed people. This goal will be met if we can sample controls from the source population such that the ratio of the number of exposed controls (B_1) to the total exposed experience of the source population is the same as the ratio of the number of unexposed controls (B_0) to the unexposed experience of the source population, apart from sampling error. For most purposes, this goal need only be followed within strata of factors that will be used for stratification in the analysis, such as factors used for restriction or matching (Chapters 11, 15, 16, and 21).

Using person-time to illustrate, the goal requires that B_1 has the same ratio to the amount of exposed person-time (T_1) as B_0 has to the amount of unexposed person-time (T_0), apart from sampling error:

$$\frac{B_1}{T_1} = \frac{B_0}{T_0}$$

Here B_1/T_1 and B_0/T_0 are the control sampling rates—that is, the number of controls selected per unit of person-time. Suppose that A_1 exposed cases and A_0 unexposed cases occur over the study period. The exposed and unexposed rates are then

$$I_1 = \frac{A_1}{T_1} \qquad \text{and} \qquad I_0 = \frac{A_0}{T_0}$$

We can use the frequencies of exposed and unexposed controls as substitutes for the actual denominators of the rates to obtain exposure-specific case-control ratios, or *pseudo-rates:*

$$\text{Pseudo-rate}_1 = \frac{A_1}{B_1}$$

and

$$\text{Pseudo-rate}_0 = \frac{A_0}{B_0}$$

These pseudo-rates have no epidemiologic interpretation by themselves. Suppose, however, that the control sampling rates B_1/T_1 and B_0/T_0 are equal to the same value r, as would be expected if controls are selected independently of exposure. If this common sampling rate r is known, the actual incidence rates can be calculated by simple algebra because, apart from sampling error, B_1/r should equal the amount of exposed person-time in the source population and B_0/r should equal the amount of unexposed person-time in the source population: $B_1/r = B_1/(B_1/T_1) = T_1$ and $B_0/r = B_0/(B_0/T_0) = T_0$. To get the incidence rates, we need only multiply each pseudo-rate by the common sampling rate, r.

If the common sampling rate is not known, which is often the case, we can still compare the sizes of the pseudo-rates by division. Specifically, if we divide the pseudo-rate for exposed by the pseudo-rate for unexposed, we obtain

$$\frac{\text{Pseudo-rate}_1}{\text{Pseudo-rate}_0} = \frac{A_1/B_1}{A_0/B_0} = \frac{A_1/[(B_1/T_1)T_1]}{A_0/[(B_0/T_0)T_0]} = \frac{A_1/(r \cdot T_1)}{A_0/(r \cdot T_0)} = \frac{A_1/T_1}{A_0/T_0}$$

In other words, the ratio of the pseudo-rates for the exposed and unexposed is an estimate of the ratio of the incidence rates in the source population, provided that the control sampling rate is independent of exposure. Thus, using the case-control study design, one can estimate the incidence rate ratio in a population without obtaining information on every subject in the population. Similar derivations in the following section on variants of case-control designs show that one can estimate the risk ratio by sampling controls from those at risk for disease at the beginning of the follow-up period (case-cohort design) and that one can estimate the incidence odds ratio by sampling controls from the noncases at the end of the follow-up period (cumulative case-control design). With these designs, the pseudo-frequencies correspond to the incidence proportions and incidence odds, respectively, multiplied by common sampling rates.

There is a statistical penalty for using a sample of the denominators rather than measuring the person-time experience for the entire source population: The precision of the estimates of the incidence rate ratio from a case-control study is less than the precision from a cohort study of the entire population that gave rise to the cases (the source population). Nevertheless, the loss of precision that stems from sampling controls will be small if the number of controls selected per case is large (usually four or more). Furthermore, the loss is balanced by the cost savings of not having to obtain information on everyone in the source population. The cost savings might allow the epidemiologist to enlarge the source population and so obtain more cases, resulting in a better overall estimate of the incidence-rate ratio, statistically and otherwise, than would be possible using the same expenditures to conduct a cohort study.

The ratio of the two pseudo-rates in a case-control study is usually written as A_1B_0/A_0B_1 and is sometimes called the *cross-product ratio.* The cross-product ratio in a case-control study can be viewed as the ratio of cases to controls among the exposed subjects (A_1/B_1), divided by the ratio of cases to controls among the unexposed subjects (A_0/B_0). This ratio can also be viewed as the odds of being exposed among cases (A_1/A_0) divided by the odds of being exposed among controls (B_1/B_0), in which case it is termed the *exposure odds ratio.* While either interpretation will give the same result, viewing this odds ratio as the ratio of case-control ratios shows more directly how the control group substitutes for the denominator information in a cohort study and how the ratio of pseudo-frequencies gives the same result as the ratio of the incidence rates, incidence proportion, or incidence odds in the source population, if sampling is independent of exposure.

One point that we wish to emphasize is that *nowhere* in the preceding discussion did we have to assume that the disease under study is "rare." In general, the rare-disease assumption is *not* needed in case-control studies. Just as for cohort studies, however, neither the incidence odds ratio nor the rate ratio should be expected to be a good approximation to the risk ratio or to be collapsible across strata of a risk factor (even if the factor is not a confounder) unless the incidence proportion is less than about 0.1 for every combination of the exposure and the factor (Chapter 4).

DEFINING THE SOURCE POPULATION

If the cases are a representative sample of all cases in a precisely defined and identified population and the controls are sampled directly from this source population, the study is said to be *population-based* or a *primary* base study. For a population-based case-control study, random sampling of controls may be feasible if a population registry exists or can be compiled. When random sampling from the source population of cases is feasible, it is usually the most desirable option.

Random sampling of controls does not necessarily mean that every person should have an equal probability of being selected to be a control. As explained earlier, if the aim is to estimate the incidence-rate ratio, then we would employ longitudinal (density) sampling, in which a person's control selection probability is proportional to the person's time at risk. For example, in a case-control study nested within an occupational cohort, workers on an employee roster will have been followed for varying lengths of time, and a random sampling scheme should reflect this varying time to estimate the incidence-rate ratio.

When it is not possible to identify the source population explicitly, simple random sampling is not feasible and other methods of control selection must be used. Such studies are sometimes called studies of *secondary* bases, because the source population is identified secondarily to the definition of a case-finding mechanism. A secondary source population or "secondary base" is therefore a source population that is defined from (secondary to) a given case series.

Consider a case-control study in which the cases are patients treated for severe psoriasis at the Mayo Clinic. These patients come to the Mayo Clinic from all corners of the world. What is the specific source population that gives rise to these cases? To answer this question, we would have to know exactly who would go to the Mayo Clinic if he or she had severe psoriasis. We cannot enumerate this source population, because many people in it do not know themselves that they would go to the Mayo Clinic for severe psoriasis, unless they actually developed severe psoriasis. This secondary source might be defined as a population spread around the world that constitutes those people who would go to the Mayo Clinic if they developed severe psoriasis. It is this secondary source from which the control series for the study would ideally be drawn. The challenge to the investigator is to apply eligibility criteria to the cases and controls so that there is good correspondence between the controls and this source population. For example, cases of severe psoriasis and controls might be restricted to those in counties within a certain distance of the Mayo Clinic, so that at least a geographic correspondence between the controls and the secondary source population could be assured. This restriction, however, might leave very few cases for study.

Unfortunately, the concept of a secondary base is often tenuously connected to underlying realities, and it can be highly ambiguous. For the psoriasis example, whether a person would go to the Mayo Clinic depends on many factors that vary over time, such as whether the person is encouraged to go by his regular physician and whether the person can afford to go. It is not clear, then, how or even whether one could precisely define, let alone sample from, the secondary base, and thus it is not clear that one could ensure that controls were members of the base at the time of sampling. We therefore prefer to conceptualize and conduct case-control studies as starting with a well-defined source population and then identify and recruit cases and controls to represent the disease and exposure experience of that population. When one instead takes a case series as a starting point, it is incumbent upon the investigator to demonstrate that a source population can be operationally defined to allow the study to be recast and evaluated relative to this source. Similar considerations apply when one takes a control series as a starting point, as is sometimes done (Greenland, 1985a).

CASE SELECTION

Ideally, case selection will amount to a direct sampling of cases within a source population. Therefore, apart from random sampling, all people in the source population who develop the disease of interest are presumed to be included as cases in the case-control study. It is not always necessary, however, to include all cases from the source population. Cases, like controls, can be randomly sampled for inclusion in the case-control study, so long as this sampling is independent of the exposure under study within strata of factors that will be used for stratification in the analysis. To see this, suppose we take only a fraction, f, of all cases. If this fraction is constant across exposure, and A_1 exposed cases and A_0 unexposed cases occur in the source population, then, apart from sampling error, the study odds ratio will be

$$\frac{A_1/B_1}{A_0/B_0} = \frac{fA_1/(r \cdot T_1)}{fA_0/(r \cdot T_0)} = \frac{A_1/T_1}{A_0/T_0}$$

as before. Of course, if fewer than all cases are sampled ($f < 1$), the study precision will be lower in proportion to f.

The cases identified in a single clinic or treated by a single medical practitioner are possible case series for case-control studies. The corresponding source population for the cases treated in a clinic is all people who would attend that clinic and be recorded with the diagnosis of interest if they had the disease in question. It is important to specify "if they had the disease in question" because clinics serve different populations for different diseases, depending on referral patterns and the reputation of the clinic in specific specialty areas. As noted above, without a precisely identified source population, it may be difficult or impossible to select controls in an unbiased fashion.

CONTROL SELECTION

The definition of the source population determines the population from which controls are sampled. Ideally, selection will involve direct sampling of controls from the source population. Based on the

principles explained earlier regarding the role of the control series, many general rules for control selection can be formulated. Two basic rules are that:

1. Controls should be selected from the same population—the source population—that gives rise to the study cases. If this rule cannot be followed, there needs to be solid evidence that the population supplying controls has an exposure distribution identical to that of the population that is the source of cases, which is a very stringent demand that is rarely demonstrable.
2. Within strata of factors that will be used for stratification in the analysis, controls should be selected independently of their exposure status, in that the sampling rate for controls (r in the previous discussion) should not vary with exposure.

If these rules and the corresponding case rules are met, then the ratio of pseudo-frequencies will, apart from sampling error, equal the ratio of the corresponding measure of disease frequency in the source population. If the sampling rate is known, then the actual measures of disease frequency can also be calculated. (If the sampling rates differ for exposed and unexposed cases or controls, but are known, the measures of disease frequency and their ratios can still be calculated using special correction formulas; see Chapters 15 and 19.) For a more detailed discussion of the principles of control selection in case-control studies, see Wacholder et al. (1992a, 1992b, 1992c).

When one wishes controls to represent person-time, sampling of the person-time should be constant across exposure levels. This requirement implies that the sampling *probability* of any person as a control should be proportional to the amount of person-time that person spends at risk of disease in the source population. For example, if in the source population one person contributes twice as much person-time during the study period as another person, the first person should have twice the probability of the second of being selected as a control. This difference in probability of selection is automatically induced by sampling controls at a steady rate per unit time over the period in which cases are sampled (*longitudinal* or *density* sampling), rather than by sampling all controls at a point in time (such as the start or end of the follow-up period). With longitudinal sampling of controls, a population member present for twice as long as another will have twice the chance of being selected.

If the objective of the study is to estimate a risk or rate ratio, it should be possible for a person to be selected as a control and yet remain eligible to become a case, so that person might appear in the study as both a control and a case. This possibility may sound paradoxical or wrong, but it is nevertheless correct. It corresponds to the fact that in a cohort study, a case contributes to both the numerator and the denominator of the estimated incidence.

Suppose the follow-up period spans 3 years, and a person free of disease in year 1 is selected as a potential control at year 1. This person should in principle remain eligible to become a case. Suppose this control now develops the disease at year 2 and now becomes a case in the study. How should such a person be treated in the analysis? Because the person did develop disease during the study period, many investigators would count the person as a case but not as a control. If the objective is to have the case-control odds ratio estimate the incidence odds ratio, then this decision would be appropriate. Recall, however, that if a follow-up study were being conducted, each person who develops disease would contribute not only to the numerator of the disease risk or rate but also to the persons or person-time tallied in the denominator. We want the control group to provide estimates of the relative size of the denominators of the incidence proportions or incidence rates for the compared groups. These denominators include all people who later become cases. Therefore, each case in a case-control study should be eligible to be a control before the time of disease onset, each control should be eligible to become a case as of the time of selection as a control, and a person selected as a control who later does develop the disease and is selected as a case should be included in the study both as a control and as a case (Sheehe, 1962; Miettinen, 1976a; Greenland and Thomas, 1982; Lubin and Gail, 1984; Robins et al., 1986a). If the controls are intended to represent person time and are selected longitudinally, similar arguments show that a person selected as a control should remain eligible to be selected as a control again, and thus might be included in the analysis repeatedly as a control (Lubin and Gail, 1984; Robins et al., 1986a).

COMMON FALLACIES IN CONTROL SELECTION

In cohort studies, the study population is restricted to people at risk for the disease. Some authors have viewed case-control studies as if they were cohort studies done backwards, even going so far as to describe them as "trohoc" studies (Feinstein, 1973). Under this view, the argument was advanced that case-control studies ought to be restricted to those at risk for exposure (i.e., those with exposure opportunity). Excluding sterile women from a case-control study of an adverse effect of oral contraceptives and matching for duration of employment in an occupational study are examples of attempts to control for exposure opportunity. If the factor used for restriction (e.g., sterility) is unrelated to the disease, it will not be a confounder, and hence the restriction will yield no benefit to the validity of the estimate of effect. Furthermore, if the restriction reduces the study size, the precision of the estimate of effect will be reduced (Poole, 1986).

Another principle sometimes used in cohort studies is that the study cohort should be "clean" at start of follow-up, including only people who have never had the disease. Misapplying this principle to case-control design suggests that the control group ought to be "clean," including only people who are healthy, for example. Illness arising after the start of the follow-up period is not reason to exclude subjects from a cohort analysis, and such exclusion can lead to bias; similarly controls with illness that arose after exposure should not be removed from the control series. Nonetheless, in studies of the relation between cigarette smoking and colorectal cancer, certain authors recommended that the control group should exclude people with colon polyps, because colon polyps are associated with smoking and are precursors of colorectal cancer (Terry and Neugut, 1998). Such an exclusion actually reduces the prevalence of the exposure in the controls below that in the source population of cases and hence biases the effect estimates upward (Poole, 1999).

SOURCES FOR CONTROL SERIES

The following methods for control sampling apply when the source population cannot be explicitly enumerated, so random selection is not possible. All of these methods should only be implemented subject to the reservations about secondary bases described earlier.

Neighborhood Controls

If the source population cannot be enumerated, it may be possible to select controls through sampling of residences. This method is not straightforward. Usually, a geographic roster of residences is not available, so a scheme must be devised to sample residences without enumerating them all. For convenience, investigators may sample controls who are individually matched to cases from the same neighborhood. That is, after a case is identified, one or more controls residing in the same neighborhood as that case are identified and recruited into the study. If neighborhood is related to exposure, the matching should be taken into account in the analysis (see Chapter 16).

Neighborhood controls are often used when the cases are recruited from a convenient source, such as a clinic or hospital. Such usage can introduce bias, however, for the neighbors selected as controls may not be in the source population of the cases. For example, if the cases are from a particular hospital, neighborhood controls may include people who would not have been treated at the same hospital had they developed the disease. If being treated at the hospital from which cases are identified is related to the exposure under study, then using neighborhood controls would introduce a bias. As an extreme example, suppose the hospital in question were a U.S. Veterans Administration hospital. Patients at these hospitals tend to differ from their neighbors in many ways. One obvious way is in regard to service history. Most patients at Veterans Administration hospitals have served in the U.S. military, whereas only a minority of their neighbors will have done so. This difference in life history can lead to differences in exposure histories (e.g., exposures associated with combat or weapons handling). For any given study, the suitability of using neighborhood controls needs to be evaluated with regard to the study variables on which the research focuses.

Random-Digit Dialing

Sampling of households based on random selection of telephone numbers is intended to simulate sampling randomly from the source population. *Random-digit dialing*, as this method has been called (Waksberg, 1978), offers the advantage of approaching all households in a designated area,

even those with unlisted telephone numbers, through a simple telephone call. The method requires considerable attention to details, however, and carries no guarantee of unbiased selection.

First, case eligibility should include residence in a house that has a telephone, so that cases and controls come from the same source population. Second, even if the investigator can implement a sampling method so that every telephone has the same probability of being called, there will not necessarily be the same probability of contacting each eligible control subject, because households vary in the number of people who reside in them, the amount of time someone is at home, and the number of operating phones. Third, making contact with a household may require many calls at various times of day and various days of the week, demanding considerable labor; many dozens of telephone calls may be required to obtain a control subject meeting specific eligibility characteristics (Wacholder et al., 1992b). Fourth, some households use answering machines, voicemail, or caller identification to screen calls and may not answer or return unsolicited calls. Fifth, the substitution of mobile telephones for land lines by some households further undermines the assumption that population members can be selected randomly by random-digit dialing. Finally, it may be impossible to distinguish accurately business from residential telephone numbers, a distinction required to calculate the proportion of nonresponders.

Random-digit-dialing controls are usually matched to cases on area code (in the United States, the first three digits of the telephone number) and exchange (the three digits following the area code). In the past, area code and prefix were related to residence location and telephone type (land line or mobile service). Thus, if geographic location or participation in mobile telephone plans was likely related to exposure, then the matching should be taken into account in the analysis. More recently, telephone companies in the United States have assigned overlaying area codes and have allowed subscribers to retain their telephone number when they move within the region, so the correspondence between assigned telephone numbers and geographic location has diminished.

Hospital- or Clinic-Based Controls

As noted above, the source population for hospital- or clinic-based case-control studies is not often identifiable, because it represents a group of people who would be treated in a given clinic or hospital if they developed the disease in question. In such situations, a random sample of the general population will not necessarily correspond to a random sample of the source population. If the hospitals or clinics that provide the cases for the study treat only a small proportion of cases in the geographic area, then referral patterns to the hospital or clinic are important to take into account in the sampling of controls. For these studies, a control series comprising patients from the same hospitals or clinics as the cases may provide a less biased estimate of effect than general-population controls (such as those obtained from case neighborhoods or by random-digit dialing). The source population does not correspond to the population of the geographic area, but only to the people who would seek treatment at the hospital or clinic were they to develop the disease under study. Although the latter population may be difficult or impossible to enumerate or even define very clearly, it seems reasonable to expect that other hospital or clinic patients will represent this source population better than general-population controls. The major problem with any nonrandom sampling of controls is the possibility that they are not selected independently of exposure in the source population. Patients who are hospitalized with other diseases, for example, may be unrepresentative of the exposure distribution in the source population, either because exposure is associated with hospitalization, or because the exposure is associated with the other diseases, or both. For example, suppose the study aims to evaluate the relation between tobacco smoking and leukemia using hospitalized cases. If controls are people who are hospitalized with other conditions, many of them will have been hospitalized for conditions associated with smoking. A variety of other cancers, as well as cardiovascular diseases and respiratory diseases, are related to smoking. Thus, a control series of people hospitalized for diseases other than leukemia would include a higher proportion of smokers than would the source population of the leukemia cases.

Limiting the diagnoses for controls to conditions for which there is no prior indication of an association with the exposure improves the control series. For example, in a study of smoking and hospitalized leukemia cases, one could exclude from the control series anyone who was hospitalized with a disease known to be related to smoking. Such an exclusion policy may exclude most of the potential controls, because cardiovascular disease by itself would represent a large proportion of hospitalized patients. Nevertheless, even a few common diagnostic categories should suffice to

find enough control subjects, so that the exclusions will not harm the study by limiting the size of the control series. Indeed, in limiting the scope of eligibility criteria, it is reasonable to exclude categories of potential controls even on the suspicion that a given category might be related to the exposure. If wrong, the cost of the exclusion is that the control series becomes more homogeneous with respect to diagnosis and perhaps a little smaller. But if right, then the exclusion is important to the ultimate validity of the study.

On the other hand, an investigator can rarely be sure that an exposure is not related to a disease or to hospitalization for a specific diagnosis. Consequently, it would be imprudent to use only a single diagnostic category as a source of controls. Using a variety of diagnoses has the advantage of potentially diluting the biasing effects of including a specific diagnostic group that is related to the exposure, and allows examination of the effect of excluding certain diagnoses.

Excluding a diagnostic category from the list of eligibility criteria for identifying controls is intended simply to improve the representativeness of the control series with respect to the source population. Such an exclusion criterion does not imply that there should be exclusions based on disease history (Lubin and Hartge, 1984). For example, in a case-control study of smoking and hospitalized leukemia patients, one might use hospitalized controls but exclude any who are hospitalized because of cardiovascular disease. This exclusion criterion for controls does not imply that leukemia cases who have had cardiovascular disease should be excluded; only if the cardiovascular disease was a cause of the hospitalization should the case be excluded. For controls, the exclusion criterion should apply only to the cause of the hospitalization used to identify the study subject. A person who was hospitalized because of a traumatic injury and who is thus eligible to be a control would not be excluded if he or she had previously been hospitalized for cardiovascular disease. The source population includes people who have had cardiovascular disease, and they should be included in the control series. Excluding such people would lead to an underrepresentation of smoking relative to the source population and produce an upward bias in the effect estimates.

If exposure directly affects hospitalization (for example, if the decision to hospitalize is in part based on exposure history), the resulting bias cannot be remedied without knowing the hospitalization rates, even if the exposure is unrelated to the study disease or the control diseases. This problem was in fact one of the first problems of hospital-based studies to receive detailed analysis (Berkson, 1946), and is often called Berksonian bias; it is discussed further under the topics of selection bias (Chapter 9) and collider bias (Chapter 12).

Other Diseases

In many settings, especially in populations with established disease registries or insurance-claims databases, it may be most convenient to choose controls from people who are diagnosed with other diseases. The considerations needed for valid control selection from other diagnoses parallel those just discussed for hospital controls. It is essential to exclude any diagnoses known or suspected to be related to exposure, and better still to include only diagnoses for which there is some evidence indicating that they are unrelated to exposure. These exclusion and inclusion criteria apply only to the diagnosis that brought the person into the registry or database from which controls are selected. The history of an exposure-related disease should not be a basis for exclusion. If, however, the exposure directly affects the chance of entering the registry or database, the study will be subject to the Berksonian bias mentioned earlier for hospital studies.

Friend Controls

Choosing friends of cases as controls, like using neighborhood controls, is a design that inherently uses individual matching and needs to be evaluated with regard to the advantages and disadvantages of such matching (discussed in Chapter 11).

Aside from the complications of individual matching, there are further concerns stemming from use of friend controls. First, being named as a friend by the case may be related to the exposure status of the potential control (Flanders and Austin, 1986). For example, cases might preferentially name as friends their acquaintances with whom they engage in specific activities that might relate to the exposure. Physical activity, alcoholic beverage consumption, and sun exposure are examples of such exposures. People who are more reclusive may be less likely to be named as friends, so their exposure patterns will be underrepresented among a control series of friends. Exposures more common to extroverted people may become overrepresented among friend controls. This type of

bias was suspected in a study of insulin-dependent diabetes mellitus in which the parents of cases identified the controls. The cases had fewer friends than controls, had more learning problems, and were more likely to dislike school. Using friend controls could explain these findings (Siemiatycki, 1989).

A second problem is that, unlike other methods of control selection, choosing friends as controls cedes much of the decision making about the choice of control subjects to the cases or their proxies (e.g., parents). The investigator who uses friend controls will usually ask for a list of friends and choose randomly from the list, but for the creation of the list, the investigator is completely dependent on the cases or their proxies. This dependence adds a potential source of bias to the use of friend controls that does not exist for other sources of controls.

A third problem is that using friend controls can introduce a bias that stems from the overlapping nature of friendship groups (Austin et al., 1989; Robins and Pike, 1990). The problem arises because different cases name groups of friends that are not mutually exclusive. As a result, people with many friends become overrepresented in the control series, and any exposures associated with such people become overrepresented as well (see Chapter 11).

In principle, matching categories should form a mutually exclusive and collectively exhaustive partition with respect to all factors, such as neighborhood and age. For example, if matching on age, bias due to overlapping matching groups can arise from *caliper matching*, a term that refers to choosing controls who have a value for the matching factor within a specified range of the case's value. Thus, if the case is 69 years old, one might choose controls who are within 2 years of age 69. Overlap bias can be avoided if one uses nonoverlapping age categories for matching. Thus, if the case is 69 years old, one might choose controls from within the age category 65 to 69 years. In practice, however, bias due to overlapping age and neighborhood categories is probably minor (Robins and Pike, 1990).

Dead Controls

A dead control cannot be a member of the source population for cases, because death precludes the occurrence of any new disease. Suppose, however, that the cases are dead. Does the need for comparability argue in favor of using dead controls? Although certain types of comparability are important, choosing dead controls will misrepresent the exposure distribution in the source population if the exposure causes or prevents death in a substantial proportion of people or if it is associated with an uncontrolled factor that does. If interviews are needed and some cases are dead, it will be necessary to use proxy respondents for the dead cases. To enhance comparability of information while avoiding the problems of taking dead controls, proxy respondents can also be used for those live controls matched to dead cases (Wacholder et al., 1992b). The advantage of comparable information for cases and controls is often overstated, however, as will be addressed later. The main justification for using dead controls is convenience, such as in studies based entirely on deaths (see the discussion of proportional mortality studies below and in Chapter 6).

OTHER CONSIDERATIONS FOR SUBJECT SELECTION

Representativeness

Some textbooks have stressed the need for representativeness in the selection of cases and controls. The advice has been that cases should be representative of all people with the disease and that controls should be representative of the entire nondiseased population. Such advice can be misleading. A case-control study may be restricted to any type of case that may be of interest: female cases, old cases, severely ill cases, cases that died soon after disease onset, mild cases, cases from Philadelphia, cases among factory workers, and so on. In none of these examples would the cases be representative of all people with the disease, yet perfectly valid case-control studies are possible in each one (Cole, 1979). The definition of a case can be quite specific as long as it has a sound rationale. The main concern is clear delineation of the population that gave rise to the cases.

Ordinarily, controls should represent the source population for cases (within categories of stratification variables), rather than the entire nondiseased population. The latter may differ vastly from the source population for the cases by age, race, sex (e.g., if the cases come from a Veterans Administration hospital), socioeconomic status, occupation, and so on—including the exposure of interest.

One of the reasons for emphasizing the similarities rather than the differences between cohort and case-control studies is that numerous principles apply to both types of study but are more evident in the context of cohort studies. In particular, many principles relating to subject selection apply identically to both types of study. For example, it is widely appreciated that cohort studies can be based on special cohorts rather than on the general population. It follows that case-control studies can be conducted by sampling cases and controls from within those special cohorts. The resulting controls should represent the distribution of exposure across those cohorts, rather than the general population, reflecting the more general rule that controls should represent the source population of the cases in the study, not the general population.

Comparability of Information Accuracy

Some authors have recommended that information obtained about cases and controls should be of comparable or equal accuracy, to ensure nondifferentiality (equal distribution) of measurement errors (Miettinen, 1985a; Wacholder et al., 1992a; MacMahon and Trichopoulos, 1996). The rationale for this principle is the notion that nondifferential measurement error biases the observed association toward the null, and so will not generate a spurious association, and that bias in studies with nondifferential error is more predictable than in studies with differential error.

The comparability-of-information (equal-accuracy) principle is often used to guide selection of controls and collection of data. For example, it is the basis for using proxy respondents instead of direct interviews for living controls whenever case information is obtained from proxy respondents. In most settings, however, the arguments for the principle are logically inadequate. One problem, discussed at length in Chapter 9, is that nondifferentiality of exposure measurement error is far from sufficient to guarantee that bias will be toward the null. Such guarantees require that the exposure errors also be *independent* of errors in other variables, including disease and confounders (Chavance et al., 1992; Kristensen, 1992), a condition that is not always plausible (Lash and Fink, 2003b). For example, it seems likely that people who conceal heavy alcohol use will also tend to understate other socially disapproved behaviors such as heavy smoking, illicit drug use, and so on.

Another problem is that the efforts to ensure equal accuracy of exposure information will also tend to produce equal accuracy of information on other variables. The direction of overall bias produced by the resulting nondifferential errors in confounders and effect modifiers can be larger than the bias produced by differential error from unequal accuracy of exposure information from cases and controls (Greenland, 1980; Brenner, 1993; Marshall and Hastrup, 1996; Marshall et al., 1999; Fewell et al., 2007). In addition, unless the exposure is binary, even independent nondifferential error in exposure measurement is not guaranteed to produce bias toward the null (Dosemeci et al., 1990). Finally, even when the bias produced by forcing equal measurement accuracy is toward the null, there is no guarantee that the bias is less than the bias that would have resulted from using a measurement with differential error (Greenland and Robins, 1985a; Drews and Greenland, 1990; Wacholder et al., 1992a). For example, in a study that used proxy respondents for cases, use of proxy respondents for the controls might lead to greater bias than use of direct interviews with controls, even if the latter results in greater accuracy of control measurements.

The comparability-of-information (equal accuracy) principle is therefore applicable only under very limited conditions. In particular, it would seem to be useful only when confounders and effect modifiers are measured with negligible error and when measurement error is reduced by using equally accurate sources of information. Otherwise, the bias from forcing cases and controls to have equal measurement accuracy may be as unpredictable as the effect of not doing so and risking differential error (unequal accuracy).

Number of Control Groups

Situations arise in which the investigator may face a choice between two or more possible control groups. Usually, there will be advantages for one group that are missing in the other, and vice versa. Consider, for example, a case-control study based on a hospitalized series of cases. Because they are hospitalized, hospital controls would be unrepresentative of the source population to the extent that exposure is related to hospitalization for the control conditions. Neighborhood controls would not suffer this problem, but might be unrepresentative of persons who would go to the hospital if they had the study disease. So which control group is better? In such situations, some

have argued that more than one control group should be used, in an attempt to address the biases from each group (Ibrahim and Spitzer, 1979). For example, Gutensohn et al. (1975), in a case-control study of Hodgkin disease, used a control group of spouses to control for environmental influences during adult life but also used a control group of siblings to control for childhood environment and sex. Both control groups are attempting to represent the same source population of cases, but have different vulnerabilities to selection biases and match on different potential confounders.

Use of multiple control groups may involve considerable labor, so is more the exception than the rule in case-control research. Often, one available control source is superior to all practical alternatives. In such settings, effort should not be wasted on collecting controls from sources likely to be biased. Interpretation of the results will also be more complicated unless the different control groups yield similar results. If the two groups produced different results, one would face the problem of explaining the differences and attempting to infer which estimate was more valid. Logically, then, the value of using more than one control group is quite limited. The control groups can and should be compared, but a lack of difference between the groups shows only that both groups incorporate similar net bias. A difference shows only that at least one is biased, but does not indicate which is best or which is worst. Only external information could help evaluate the likely extent of bias in the estimates from different control groups, and that same external information might have favored selection of only one of the control groups at the design stage of the study.

Timing of Classification and Diagnosis

Chapter 7 discussed at length some basic principles for classifying persons, cases, and person-time units in cohort studies according to exposure status. The same principles apply to cases and controls in case-control studies. If the controls are intended to represent person-time (rather than persons) in the source population, one should apply principles for classifying person-time to the classification of controls. In particular, principles of person-time classification lead to the rule that controls should be classified by their exposure status as of their selection time. Exposures accrued after that time should be ignored. The rule necessitates that information (such as exposure history) be obtained in a manner that allows one to ignore exposures accrued after the selection time. In a similar manner, cases should be classified as of time of diagnosis or disease onset, accounting for any built-in lag periods or induction-period hypotheses. Determining the time of diagnosis or disease onset can involve all the problems and ambiguities discussed in the previous chapter for cohort studies and needs to be resolved by study protocol before classifications can be made.

As an example, consider a study of alcohol use and laryngeal cancer that also examined smoking as a confounder and possible effect modifier, used interviewer-administered questionnaires to collect data, and used neighborhood controls. To examine the effect of alcohol and smoking while assuming a 1-year lag period (a 1-year minimum induction time), the questionnaire would have to allow determination of drinking and smoking habits up to 1 year before diagnosis (for cases) or selection (for controls).

Selection time need not refer to the investigator's identification of the control, but instead may refer to an event analogous to the occurrence time for the case. For example, the selection time for controls who are cases of other diseases can be taken as time of diagnosis for that disease; the selection time of hospital controls might be taken as time of hospitalization. For other types of controls, there may be no such natural event analogous to the case diagnosis time, and the actual time of selection will have to be used.

In most studies, selection time will precede the time data are gathered. For example, in interview-based studies, controls may be identified and then a delay of weeks or months may occur before the interview is conducted. To avoid complicating the interview questions, this distinction is often ignored and controls are questioned about habits in periods dating back from the interview.

VARIANTS OF THE CASE-CONTROL DESIGN

NESTED CASE-CONTROL STUDIES

Epidemiologists sometimes refer to specific case-control studies as *nested* case-control studies when the population within which the study is conducted is a fully enumerated cohort, which allows formal

random sampling of cases and controls to be carried out. The term is usually used in reference to a case-control study conducted within a cohort study, in which further information (perhaps from expensive tests) is obtained on most or all cases, but for economy is obtained from only a fraction of the remaining cohort members (the controls). Nonetheless, many population-based case-control studies can be thought of as nested within an enumerated source population. For example, when there is a population-based disease registry and a census enumeration of the population served by the registry, it may be possible to use the census data to sample controls randomly.

CASE-COHORT STUDIES

The *case-cohort study* is a case-control study in which the source population is a cohort and (within sampling or matching strata) every person in this cohort has an equal chance of being included in the study as a control, regardless of how much time that person has contributed to the person-time experience of the cohort or whether the person developed the study disease. This design is a logical way to conduct a case-control study when the effect measure of interest is the ratio of incidence proportions rather than a rate ratio, as is common in perinatal studies. The average risk (or incidence proportion) of falling ill during a specified period may be written

$$R_1 = \frac{A_1}{N_1}$$

for the exposed subcohort and

$$R_0 = \frac{A_0}{N_0}$$

for the unexposed subcohort, where R_1 and R_0 are the incidence proportions among the exposed and unexposed, respectively, and N_1 and N_0 are the initial sizes of the exposed and unexposed subcohorts. (This discussion applies equally well to exposure variables with several levels, but for simplicity we will consider only a dichotomous exposure.) Controls should be selected such that the exposure distribution among them will estimate without bias the exposure distribution in the source population. In a case-cohort study, the distribution we wish to estimate is among the $N_1 + N_0$ cohort members, not among their person-time experience (Thomas, 1972; Kupper et al., 1975; Miettinen, 1982a).

The objective is to select controls from the source cohort such that the ratio of the number of exposed controls (B_1) to the number of exposed cohort members (N_1) is the same as the ratio of the number of unexposed controls (B_0) to the number of unexposed cohort members (N_0), apart from sampling error:

$$\frac{B_1}{N_1} = \frac{B_0}{N_0}$$

Here, B_1/N_1 and B_0/N_0 are the control sampling fractions (the number of controls selected per cohort member). Apart from random error, these sampling fractions will be equal if controls have been selected independently of exposure.

We can use the frequencies of exposed and unexposed controls as substitutes for the actual denominators of the incidence proportions to obtain "pseudo-risks":

$$\text{Pseudo-risk}_1 = \frac{A_1}{B_1}$$

and

$$\text{Pseudo-risk}_0 = \frac{A_0}{B_0}$$

These pseudo-risks have no epidemiologic interpretation by themselves. Suppose, however, that the control sampling fractions are equal to the same fraction, f. Then, apart from sampling error, B_1/f should equal N_1, the size of the exposed subcohort; and B_0/f should equal N_0, the size of the unexposed subcohort: $B_1/f = B_1/(B_1/N_1) = N_1$ and $B_0/f = B_0/(B_0/N_0) = N_0$. Thus, to get the

incidence proportions, we need only multiply each pseudo-risk by the common sampling fraction, f. If this fraction is not known, we can still compare the sizes of the pseudo-risks by division:

$$\frac{\text{Pseudo-risk}_1}{\text{Pseudo-risk}_0} = \frac{A_1/B_1}{A_0/B_0} = \frac{A_1/[(B_1/N_1)N_1]}{A_0/[(B_0/N_0)N_0]} = \frac{A_1/fN_1}{A_0/fN_0} = \frac{A_1/N_1}{A_0/N_0}$$

In other words, the ratio of pseudo-risks is an estimate of the ratio of incidence proportions (risk ratio) in the source cohort if control sampling is independent of exposure. Thus, using a case-cohort design, one can estimate the risk ratio in a cohort without obtaining information on every cohort member.

Thus far, we have implicitly assumed that there is no loss to follow-up or competing risks in the underlying cohort. If there are such problems, it is still possible to estimate risk or rate ratios from a case-cohort study, provided that we have data on the time spent at risk by the sampled subjects or we use certain sampling modifications (Flanders et al., 1990). These procedures require the usual assumptions for rate-ratio estimation in cohort studies, namely, that loss-to-follow-up and competing risks either are not associated with exposure or are not associated with disease risk.

An advantage of the case-cohort design is that it facilitates conduct of a set of case-control studies from a single cohort, all of which use the same control group. As a sample from the cohort at enrollment, the control group can be compared with any number of case groups. If matched controls are selected from people at risk at the time a case occurs (as in risk-set sampling, which is described later), the control series must be tailored to a specific group of cases. If common outcomes are to be studied and one wishes to use a single control group for each outcome, another sampling scheme must be used. The case-cohort approach is a good choice in such a situation.

Case-cohort designs have other advantages as well as disadvantages relative to alternative case-control designs (Wacholder, 1991). One disadvantage is that, because of the overlap of membership in the case and control groups (controls who are selected may also develop disease and enter the study as cases), one will need to select more controls in a case-cohort study than in an ordinary case-control study with the same number of cases, if one is to achieve the same amount of statistical precision. Extra controls are needed because the statistical precision of a study is strongly determined by the numbers of distinct cases and noncases. Thus, if 20% of the source cohort members will become cases, and all cases will be included in the study, one will have to select 1.25 times as many controls as cases in a case-cohort study to ensure that there will be as many controls who never become cases in the study. On average, only 80% of the controls in such a situation will remain noncases; the other 20% will become cases. Of course, if the disease is uncommon, the number of extra controls needed for a case-cohort study will be small.

DENSITY CASE-CONTROL STUDIES

Earlier, we described how case-control odds ratios will estimate rate ratios if the control series is selected so that the ratio of the person-time denominators T_1/T_0 is validly estimated by the ratio of exposed to unexposed controls B_1/B_0. That is, to estimate rate ratios, controls should be selected so that the exposure distribution among them is, apart from random error, the same as it is among the person-time in the source population or within strata of the source population. Such control selection is called *density sampling* because it provides for estimation of relations among incidence rates, which have been called *incidence densities*.

If a subject's exposure may vary over time, then a case's exposure history is evaluated up to the time the disease occurred. A control's exposure history is evaluated up to an analogous index time, usually taken as the time of sampling; exposure after the time of selection must be ignored. This rule helps ensure that the number of exposed and unexposed controls will be in proportion to the amount of exposed and unexposed person-time in the source population.

The time during which a subject is eligible to be a control should be the time in which that person is also eligible to become a case, if the disease should occur. Thus, a person in whom the disease has already developed or who has died is no longer eligible to be selected as a control. This rule corresponds to the treatment of subjects in cohort studies. Every case that is tallied in the numerator of a cohort study contributes to the denominator of the rate until the time that the person

becomes a case, when the contribution to the denominator ceases. One way to implement this rule is to choose controls from the set of people in the source population who are at risk of becoming a case at the time that the case is diagnosed. This set is sometimes referred to as the *risk set* for the case, and this type of control sampling is sometimes called *risk-set sampling.* Controls sampled in this manner are matched to the case with respect to sampling time; thus, if time is related to exposure, the resulting data should be analyzed as matched data (Greenland and Thomas, 1982). It is also possible to conduct unmatched density sampling using probability sampling methods if one knows the time interval at risk for each population member. One then selects a control by sampling members with probability proportional to time at risk and then randomly samples a time to measure exposure within the interval at risk.

As mentioned earlier, a person selected as a control who remains in the study population at risk after selection should remain eligible to be selected once again as a control. Thus, although it is unlikely in typical studies, the same person may appear in the control group two or more times. Note, however, that including the same person at different times does not necessarily lead to exposure (or confounder) information being repeated, because this information may change with time. For example, in a case-control study of an acute epidemic of intestinal illness, one might ask about food ingested within the previous day or days. If a contaminated food item was a cause of the illness for some cases, then the exposure status of a case or control chosen 5 days into the study might well differ from what it would have been 2 days into the study.

CUMULATIVE ("EPIDEMIC") CASE-CONTROL STUDIES

In some research settings, case-control studies may address a risk that ends before subject selection begins. For example, a case-control study of an epidemic of diarrheal illness after a social gathering may begin after all the potential cases have occurred (because the maximum induction time has elapsed). In such a situation, an investigator might select controls from that portion of the population that remains after eliminating the accumulated cases; that is, one selects controls from among noncases (those who remain noncases at the end of the epidemic follow-up).

Suppose that the source population is a cohort and that a fraction f of both exposed and unexposed noncases is selected to be controls. Then the ratio of pseudo-frequencies will be

$$\frac{A_1/B_1}{A_0/B_0} = \frac{A_1/f(N_1 - A_1)}{A_0/f(N_0 - A_0)} = \frac{A_1/(N_1 - A_1)}{A_0/(N_0 - A_0)}$$

which is the incidence odds ratio for the cohort. This ratio will provide a reasonable approximation to the rate ratio, provided that the proportions falling ill in each exposure group during the risk period are low, that is, less than about 20%, and that the prevalence of exposure remains reasonably steady during the study period (see Chapter 4). If the investigator prefers to estimate the risk ratio rather than the incidence rate ratio, the study odds ratio can still be used (Cornfield, 1951), but the accuracy of this approximation is only about half as good as that of the odds-ratio approximation to the rate ratio (Greenland, 1987a). The use of this approximation in the cumulative design is the primary basis for the mistaken teaching that a rare-disease assumption is needed to estimate effects from case-control studies.

Before the 1970s, the standard conceptualization of case-control studies involved the cumulative design, in which controls are selected from noncases at the end of a follow-up period. As discussed by numerous authors (Sheehe, 1962; Miettinen, 1976a; Greenland and Thomas, 1982), density designs and case-cohort designs have several advantages outside of the acute epidemic setting, including potentially much less sensitivity to bias from exposure-related loss-to-follow-up.

CASE-ONLY, CASE-SPECULAR, AND CASE-CROSSOVER STUDIES

There are a number of situations in which cases are the only subjects used to estimate or test hypotheses about effects. For example, it is sometimes possible to employ theoretical considerations to construct a prior distribution of exposure in the source population and use this distribution in place of an observed control series. Such situations arise naturally in genetic studies, in which basic

laws of inheritance may be combined with certain assumptions to derive a population or parental-specific distribution of genotypes (Self et al., 1991). It is also possible to study certain aspects of joint effects (interactions) of genetic and environmental factors without using control subjects (Khoury and Flanders, 1996); see Chapter 28 for details.

When the exposure under study is defined by proximity to an environmental source (e.g., a power line), it may be possible to construct a *specular* (hypothetical) control for each case by conducting a "thought experiment." Either the case or the exposure source is imaginarily moved to another location that would be equally likely were there no exposure effect; the case exposure level under this hypothetical configuration is then treated as the (matched) "control" exposure for the case (Zaffanella et al., 1998). When the specular control arises by examining the exposure experience of the case outside of the time in which exposure could be related to disease occurrence, the result is called a *case-crossover study*.

The classic *crossover* study is a type of experiment in which two (or more) treatments are compared, as in any experimental study. In a crossover study, however, each subject receives both treatments, with one following the other. Preferably, the order in which the two treatments are applied is randomly chosen for each subject. Enough time should be allocated between the two administrations so that the effect of each treatment can be measured and can subside before the other treatment is given. A persistent effect of the first intervention is called a *carryover effect*. A crossover study is only valid to study treatments for which effects occur within a short induction period and do not persist, i.e., carryover effects must be absent, so that the effect of the second intervention is not intermingled with the effect of the first.

The *case-crossover* study is a case-control analog of the crossover study (Maclure, 1991). For each case, one or more predisease or postdisease time periods are selected as matched "control" periods for the case. The exposure status of the case at the time of the disease onset is compared with the distribution of exposure status for that same person in the control periods. Such a comparison depends on the assumption that neither exposure nor confounders are changing over time in a systematic way.

Only a limited set of research topics are amenable to the case-crossover design. The exposure must vary over time within individuals rather than stay constant. Eye color or blood type, for example, could not be studied with a case-crossover design because both are constant. If the exposure does not vary within a person, then there is no basis for comparing exposed and unexposed time periods of risk within the person. Like the crossover study, the exposure must also have a short induction time and a transient effect; otherwise, exposures in the distant past could be the cause of a recent disease onset (a carryover effect).

Maclure (1991) used the case-crossover design to study the effect of sexual activity on incident myocardial infarction. This topic is well suited to a case-crossover design because the exposure is intermittent and is presumed to have a short induction period for the hypothesized effect. Any increase in risk for a myocardial infarction from sexual activity is presumed to be confined to a short time following the activity. A myocardial infarction is an outcome that is well suited to this type of study because it is thought to be triggered by events close in time. Other possible causes of a myocardial infarction that might be studied by a case-crossover study would be caffeine consumption, alcohol consumption, carbon monoxide exposure, drug exposures, and heavy physical exertion (Mittleman et al., 1993), all of which occur intermittently.

Each case and its control in a case-crossover study is automatically matched on all characteristics (e.g., sex and birth date) that do not change within individuals. Matched analysis of case-crossover data controls for all such fixed confounders, whether or not they are measured. Subject to special assumptions, control for measured time-varying confounders may be possible using modeling methods for matched data (see Chapter 21). It is also possible to adjust case-crossover estimates for bias due to time trends in exposure through use of longitudinal data from a nondiseased control group (case-time controls) (Suissa, 1995). Nonetheless, these trend adjustments themselves depend on additional no-confounding assumptions and may introduce bias if those assumptions are not met (Greenland, 1996b).

There are many possible variants of the case-crossover design, depending on how control time periods are selected. These variants offer trade-offs among potential for bias, inefficiency, and difficulty of analysis; see Lumley and Levy (2000), Vines and Farrington (2001), Navidi and Weinhandl (2002), and Janes et al. (2004, 2005) for further discussion.

TWO-STAGE SAMPLING

Another variant of the case-control study uses two-stage or two-phase sampling (Walker, 1982a; White, 1982b). In this type of study, the control series comprises a relatively large number of people (possibly everyone in the source population), from whom exposure information or perhaps some limited amount of information on other relevant variables is obtained. Then, for only a subsample of the controls, more detailed information is obtained on exposure or on other study variables that may need to be controlled in the analysis. More detailed information may also be limited to a subsample of cases. This two-stage approach is useful when it is relatively inexpensive to obtain the exposure information (e.g., by telephone interview), but the covariate information is more expensive to obtain (say, by laboratory analysis). It is also useful when exposure information already has been collected on the entire population (e.g., job histories for an occupational cohort), but covariate information is needed (e.g., genotype). This situation arises in cohort studies when more information is required than was gathered at baseline. As will be discussed in Chapter 15, this type of study requires special analytic methods to take full advantage of the information collected at both stages.

PROPORTIONAL MORTALITY STUDIES

Proportional mortality studies were discussed in Chapter 6, where the point was made that the validity of such studies can be improved if they are designed and analyzed as case-control studies. The cases are deaths occurring within the source population. Controls are not selected directly from the source population, which consists of living people, but are taken from other deaths within the source population. This control series is acceptable if the exposure distribution within this group is similar to that of the source population. Consequently, the control series should be restricted to categories of death that are not related to the exposure. See Chapter 6 for a more detailed discussion.

CASE-CONTROL STUDIES WITH PREVALENT CASES

Case-control studies are sometimes based on prevalent cases rather than incident cases. When it is impractical to include only incident cases, it may still be possible to select existing cases of illness at a point in time. If the prevalence odds ratio in the population is equal to the incidence-rate ratio, then the odds ratio from a case-control study based on prevalent cases can unbiasedly estimate the rate ratio. As noted in Chapter 4, however, the conditions required for the prevalence odds ratio to equal the rate ratio are very strong, and a simple general relation does not exist for age-specific ratios. If exposure is associated with duration of illness or migration out of the prevalence pool, then a case-control study based on prevalent cases cannot by itself distinguish exposure effects on disease incidence from the exposure association with disease duration or migration, unless the strengths of the latter associations are known. If the size of the exposed or the unexposed population changes with time or there is migration into the prevalence pool, the prevalence odds ratio may be further removed from the rate ratio. Consequently, it is always preferable to select incident rather than prevalent cases when studying disease etiology.

As discussed in Chapter 3, prevalent cases are usually drawn in studies of congenital malformations. In such studies, cases ascertained at birth are prevalent because they have survived with the malformation from the time of its occurrence until birth. It would be etiologically more useful to ascertain all incident cases, including affected abortuses that do not survive until birth. Many of these, however, do not survive until ascertainment is feasible, and thus it is virtually inevitable that case-control studies of congenital malformations are based on prevalent cases. In this example, the source population comprises all conceptuses, and miscarriage and induced abortion represent emigration before the ascertainment date. Although an exposure will not affect duration of a malformation, it may very well affect risks of miscarriage and abortion.

Other situations in which prevalent cases are commonly used are studies of chronic conditions with ill-defined onset times and limited effects on mortality, such as obesity, Parkinson's disease, and multiple sclerosis, and studies of health services utilization.